Elastic Network Models in Biology: From Protein Mode Spectra to Chromatin Dynamics

by

She Zhang

B.S., Nanjing University of Chinese Medicine, 2014

Submitted to the Graduate Faculty of the

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This thesis or dissertation was presented

by

She Zhang

It was defended on

April 24, 2020

and approved by

Dr. Carl Kingsford, Ph.D., Herbert A. Simon Professor of Computer Science, Computational Biology Department, School of Computer Science, Carnegie Mellon University

Dr. Anne-Ruxandra Carvunis, Ph.D., Assistant Professor, Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh

Dr. David C. Whitcomb, M.D., Ph.D., Giant Eagle Foundation Professor of Cancer Genetics & Professor of Medicine, Cell Biology and Physiology, and Human Genetics, School of Medicine, University of Pittsburgh

Dissertation Director: Dr. Ivet Bahar, Ph.D., Distinguished Professor & John K. Vries Chair, Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh

Elastic Network Models in Biology:
From Protein Mode Spectra to Chromatin Dynamics

She Zhang, Ph.D.

University of Pittsburgh, 2020

Biomacromolecules perform their functions by accessing conformations energetically favored by their structure-encoded equilibrium dynamics. Elastic network model (ENM) analysis has been widely used to decompose the equilibrium dynamics of a given molecule into a spectrum of modes of motions, which separates robust, global motions from local fluctuations. The scalability and flexibility of the ENMs permit us to efficiently analyze the spectral dynamics of large systems or perform comparative analysis for large datasets of structures. I showed in this thesis how ENMs can be adapted (1) to analyze protein superfamilies that share similar tertiary structures but may differ in their sequence and functional dynamics, and (2) to analyze chromatin dynamics using contact data from Hi-C experiments, and (3) to perform a comparative analysis of genome topology across different types of cell lines. The first study showed that protein family members share conserved, highly cooperative (global) modes of motion. A low-to-intermediate frequency spectral regime was shown to have a maximal impact on the functional differentiation of families into subfamilies. The second study demonstrated the Gaussian Network Model (GNM) can accurately model chromosomal mobility and couplings between genomic loci at multiple scales: it can quantify the spatial fluctuations in the positions of gene loci, detect large genomic compartments and smaller topologically-associating domains (TADs) that undergo *en bloc* movements, and identify dynamically coupled distal regions along the chromosomes. The third study revealed close similarities between chromosomal dynamics across different cell lines on a global scale, but notable cell-specific variations in the spatial fluctuations of genomic loci. It also called attention to the role of the intrinsic spatial dynamics of chromatin as a determinant of cell differentiation. Together, these studies provide a comprehensive view of the versatility and utility of the ENMs in analyzing spatial dynamics of biomolecules, from individual proteins to the entire chromatin.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3D	Three-dimensional
AMI	Adjusted Mutual Information
ANM	Anisotropic Network Model
APBS	Architecture protein binding site
CCDD	cross-correlated distal domain
CTCF	CCCTC-binding factor
DAT	Dopamine transporter
DFS	Depth-first search
EC	Extracellular
EL	Extracellular loop
ENM	Elastic Network Model
GEO	Gene Expression Omnibus
GNM	Gaussian Network Model
HEG	Highly expressed gene
HF	High frequency
HiDeF	Hierarchical community Decoding Framework
HMG	Highly mobile gene
HMM	Histone modification marker
IC	Intracellular
IF	Inward-facing

IFc/OFo	Inward-facing closed/open
IL	Intracellular loop
kb	kilobases
LeuT	Leucine Transporter
LF	Low frequency
LTIF	Low-to-intermediate frequency
Mb	megabases
MD	Molecular Dynamics
mESC	mouse Embryonic Stem Cell
MSF	Mean-Square Fluctuation
MST	Maximum/Minimum Spanning Tree
NJ	Neighbor joining
NMA	Normal Mode Analysis
OF	Outward-facing
OFc/OFo	Outward-facing closed/open
PCA	Principle Component Analysis
PDB	Protein Data Bank
PPI	Protein-protein interaction
QPP	Quadratic placement problem
RMSD	Root-mean-square deviation
SignDy	Signature Dynamics
TAD	Topologically-Associating Domains
TFIIIC	Transcription factor for polymerase III C

TM	Transmembrane (domain/helix)
TM-score	Template modeling score
TPM	Transcripts per kilobase million
UPGMA	Unweighted pair group method with arithmetic mean
VCnorm	Vanilla Coverage normalization
VI	Variation of Information

Background

Biomolecular structures are not rigid under equilibrium conditions. Instead, they experience conformational changes ranging from small vibrations to collective movements that may involve large parts of the structure, driven by thermal fluctuations, or triggered by ligand binding and other changes in intermolecular interactions or environmental conditions. These conformational changes that are accessible to the molecule under equilibrium conditions are called equilibrium or intrinsic dynamics (Bahar, et al., 2007; Tsai, et al., 1999; Xu, et al., 2008). To a great extent, the intrinsic dynamics is determined by the way a structure is "wired", such that knowledge of only the three-dimensional (3D) structure (and thereby contacts/interactions between the atoms of a molecule) is sufficient to deduce to a good approximation how the structure fluctuates in a perturbation-free environment, and how it would respond to perturbations elicited by, for example, oligomerization or substrate binding. Therefore, the pre-existing, structure-encoded intrinsic dynamics of biomolecular systems often carry functional significance and serve as a bridge between their 3D structure and molecular function (Bahar, et al., 2015; Bahar, et al., 2010).

In recent years, there has been a surge in the number of studies that use elastic network models (ENMs) and normal mode analysis (NMA) for exploring the intrinsic dynamics of biological macromolecules (Bahar and Rader, 2005; Cui and Bahar, 2005). These studies have proven the usefulness of NMA-related methods for extracting collective modes of motions in helping (i) identify residues that play critical roles in mediating cooperative events, (ii) refine structures for docking algorithms, and (iii) steer molecular dynamics (MD) simulations to explore longer

timescales and larger length scales. Assuming that the system is stabilized by harmonic potentials, NMA provides valuable information on the equilibrium dynamics accessible to a system at different scales (Amadei, et al., 1993; Bahar, et al., 2010). Modes that describe the movement of large parts of the structure, and usually occur at the low frequency end of the mode spectrum (i.e. slow modes), are called global or essential modes, as opposed to local modes that correspond to regional fluctuations and are often associated with high frequencies (i.e. fast modes). Such a multiscale view of the space of motions can be useful for a mechanistic characterization of the movements intrinsically accessible to the molecule of interest, as it divides the complex motions into different frequency regimes and practically separates the collective, often functional, movements from the random fluctuations that occur usually on a local scale. This type of mode decomposition and its functional significance will be illustrated and validated by way of several applications in this thesis.

In addition to the physical insights that the ENMs provide into the functional dynamics of biomolecules, there are three features that make them attractive: simplicity, robustness, and scalability (Bahar, et al., 2010). In the ENMs, complex structures of biomolecules are reduced to a network of nodes and springs, where atomic details are simplified and residues or other building blocks are represented at a coarse-grained level by network nodes; inter- and intramolecular interactions driven by physiochemical forces are approximated and unified by harmonic potentials/springs. Such simplicity enables the analytical evaluation of a mode spectrum uniquely defined by the network topology for each structure modeled as an ENM. The robustness of the global motions is twofold. On one hand, the global modes derived by the ENMs are methodologically determined by the overall shape of the biomolecule and insensitive to detailed structure and energetics. On the other hand, evolutionary pressures on the molecular functions

discourage huge detrimental changes in equilibrium dynamics upon minor perturbations. This underscores the functional significance of structure-encoded dynamics, and the importance of functional global modes in the evolutionary selection of structures (Zhang, et al., 2020).

The efficiency and robustness of ENM predictions entail the most important feature for this study: scalability. The scalability of ENMs makes them particularly useful for investigating large biomolecular systems, or even further, a family of them. We show in Section 1.0 that the ENMs can be improved and developed to systematically examine the intrinsic dynamics of protein families and identify signature dynamics that is robust and unique for a given protein family, and shared by all members of that family. Additionally, the network nodes can be selected at different coarse-graining levels to accommodate different size systems, or different resolution data, and the ENM would still yield consistent results. In Section 2.0 and 3.0, we applied the ENMs to predict and characterize chromatin dynamics, a new and rising research area of interest in the field of genomics and epigenetics where experimental data are still limited to low resolutions, and we made a detailed comparison of the chromatin dynamics across several morphologically distinct cell lines and demonstrated how differences in the loci-specific spatial dynamics between different cell types relate to their differences in gene expression levels.

Much of the work presented in this thesis has been published. As the copyrights permit, some of the materials from previous publications (Bahar, et al., 2015; Ponzoni, et al., 2018; Sauerwald, et al., 2017; Zhang, et al., 2020; Zhang, et al., 2019) are reused or quoted with proper citations in the following chapters. Some of the studies were accomplished in collaboration with other people whose contributions are acknowledged where appropriate. All presented studies were conducted under the supervision and guidance of my doctoral advisor, Dr. Ivet Bahar.

1.0 Comparative Studies of Protein Dynamics Using Elastic Network Models

The relationship between protein sequence, structure, and function has been one of the most intriguing problems in molecular and structural biology (Babu, 2016; Forman-Kay and Mittag, 2013; Redfern, et al., 2008). While it is established theoretically that the tertiary structures of proteins can be well predicted solely based on their sequence, protein structures are found to be more evolutionarily conserved than the sequences. Evident structural similarities have been found among proteins with sequence identities as low as ~20%. This observation enabled a line of research in developing computational techniques for comparative modeling of protein structures *in silico*, such as homology modeling, structure classification, etc.

Many computational tools have been developed over the last couple of decades to comparatively analyze the sequential or structural differences of proteins (El-Gebali, et al., 2019; Holm and Laakso, 2016; Holm and Rosenstrom, 2010; Ilyin, et al., 2003; Knudsen and Wiuf, 2010). However, such tools are still scant and remain to be developed for investigating the differentiation of molecular functions driven by these differences. Studies in recent years have established the role of structural dynamics, also called intrinsic dynamics, in facilitating, if not driving, the interactions and function of biomolecular systems in the cell. Many biological events, including substrate recognition, binding and transport, allosteric signaling, communication and regulation, and mechanochemical responses, shortly referred to as protein actions, take advantage of the proteins' intrinsic dynamics (Bahar, et al., 2017; Bakan and Bahar, 2009). In the meantime, the evolutionary significance of global modes of motion became clear (Carnevale, et al., 2006; Hollup, et al., 2011; Maguid, et al., 2008; Maguid, et al., 2006). Computations highlighted the coupling between sequence evolution and intrinsic dynamics, and experiments demonstrated that

the changes in structure (or oligomerization state) stabilized by mutations bear close resemblance to structural changes that accommodate ligand binding (Perica, et al., 2014). Evolvability of intrinsic dynamics thus emerged as a major mechanism enabling adaptability to environmental changes, intermolecular interactions, or even mutations (Haliloglu and Bahar, 2015; Tokuriki and Tawfik, 2009). Recent work further showed that intrinsic dynamics is a major determinant of the impact of missense mutations on function, and that the inclusion of ENM-based features in a machine learning classifier improves the accuracy of pathogenicity predictions. These observations provide support for using ENM-evaluated intrinsic dynamics for analyzing the dynamics and the functional differentiation of proteins.

In the following sections, we show how ENMs can be used to extract information on the shared dynamics of members of a protein fold superfamily; and then how these procedures can be generalized and implemented as a computing pipeline for analyses alike. We applied this pipeline to large datasets of protein families with hundreds of members each, in order to understand the conservation/differentiation properties of the 3D dynamics and the contribution of collective motions in different frequency regimes to the functional mechanisms of motions.

1.1 Evolutionary Conservation of LeuT Superfamily Members

1.1.1 Introduction

Secondary active transporters translocate small molecules such as neurotransmitters, nutrients and metabolites across cellular membranes, using the energy provided by the co-transport (symport) or exchange (antiport) of ions or other solutes down their electrochemical gradients. Remarkably, several secondary active transporters, though belonging to genetically and functionally distant families, share a common architecture (or fold). Four common folds among transporters are the LeuT-, MFS-, GltPh- and NhaA-folds (Drew and Boudker, 2016; Shi, 2013). We will focus here on the LeuT-fold superfamily, which probably has the broadest representation in the Protein Data Bank (PDB, http://www.rcsb.org/) (Berman, et al., 2000) among the four folds.

The LeuT fold, first resolved for a bacterial leucine transporter (Yamashita, et al., 2005), is composed of 12 TM helices that alternate between outward-facing (OF) and inward-facing (IF) conformations during the transport cycle. The former favors the uptake of substrate from the extracellular (EC) region, and the latter its release to the intracellular (IC) region, accompanied by the cotransport of Na^+ ions, and in some cases by the antiport of other substrates/ ions (Kazmier, et al., 2017) (**Figure 1.1**). LeuT fold family members include the dopamine transporter (DAT), the multihydrophobic amino acid transporter (MhsT), the benzyl-hydantoin transporter (Mhp1), sodium/galactose transporter (vSGLT), the glycine betaine transporter (BetP), the carnitine/butyrobetaine antiporter (CaiT), and the arginine/agmatine antiporter (AdiC). An immediate question concerning the selection of a particular fold by a large number of transporters involved in different functions, and vastly differing in their sequence, is “what is special about this

fold that lend themselves to different functionalities?". What are the structural and dynamic characteristics of the LeuT fold that are exploited, or how do they adapt to different functions?

To address this fundamental question, we first examine the sequence and structure properties of LeuT superfamily members, and then proceed to their dynamics to determine a shared "signature" mobility profile that allosterically engages all TM helices. We further examine the role of structural irregularities such as helical disruptions, and that of multimerization, in the differentiation or allosteric modulation of transport activities, and determine the dynamics landscape of a large ensemble of structures sharing the LeuT-fold, which indicates the collective motions that underlie the OF → IF transition or the multimerization of LeuT-fold members. Our analysis sheds light into the ways in which these transporters achieve functional differentiation, while efficiently recruiting the same fold whose modular dynamics is exploited.

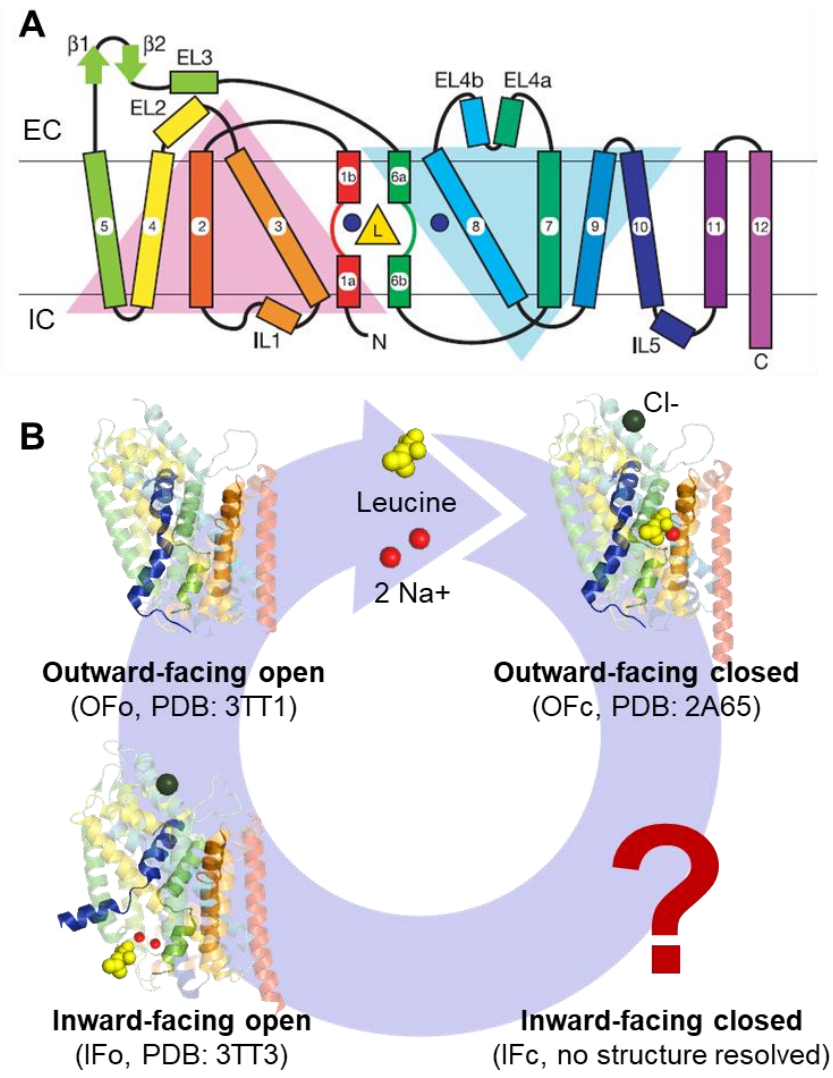

Figure 1.1 The topology and transport cycle of LeuT. (A) LeuT topology represented by its secondary structures: *rectangles*: *α*-helices; *arrows*: *β*-strands; *lines*: loops. The positions of leucine and the two sodium ions are depicted as a *yellow triangle* and *blue circles*, respectively. EC: extracellular; IC: intracellular; EL: extracellular loop; IC: intracellular loop. The transmembrane helices are indexed as TM1-12. The *pink* and *indigo* triangles in the background indicate two pseudo-symmetric inverted repeats, TM1-5 and TM6-10, commonly shared by the LeuT-fold proteins (adapted from (Yamashita, et al., 2005)). **(B)** The transport cycle of LeuT. Each state is represented by a PDB structure except for the Inward-facing closed (IFc) state for which no structure has been resolved. The structures are oriented similarily so that the *top/bottom* correspond to the EC/IC side, and the *left/rightmost helices* are TM5/12. Three TM helices are highlighted for their important role in defining the transport states: TM1 (*blue*), TM6 (*green*), TM10 (*orange*).

1.1.2 Results

We consider a set of 90 structures (104 conformers) with LeuT-fold deposited in the PDB, which belong to five functional families, listed in **Table** 1.1. The set includes the crystallographic structures resolved for eight different transporters: LeuT in different conformational states, DAT, and MhsT from the NSS family; galactose transporter (vSGLT) from the sodium/solute symporter (SSS) family; benzylhydantoin (BH) transporter Mhp1 from the nucleobase/cation symport-1 (NCS1) family; and arginine/agmatine antiporter AdiC from the amino acid/polyamine/organocation (APC) family; betaine transporter (BetP) and carnitine/betaine antiporter (CaiT) in multiple states from the betaine/choline/carnitine transporters (BCCT). Different from transporters in previous families which are typically monomeric or dimeric, BetP and CaiT in the BCCT family have been observed to robustly form trimers (Kalayil, et al., 2013; Koshy and Ziegler, 2015; Perez, et al., 2012; Ressl, et al., 2009; Schulze, et al., 2010).

1.1.2.1 Sequence differences confer specificity while maintaining the fold

The LeuT-fold (**Figure 1.1A**) is characterized by 10 TM helices, organized into two pseudo-symmetric inverted repeats, TM1-TM5 and TM6-TM10 (Schulze, et al., 2010). **Figure 1.2** displays the superposition of the transporters resolved in the OF (**panel A**) and the IF state (**panel B**), highlighting the common fold shared by the superfamily (**panel C**), as well as the distinctive packing of TM helices to expose the EC or IC vestibule in the OF and IF states, respectively. Structural alignments of the transporters listed in **Table** 1.1 reveal differences of up to 6.5 Å root-mean-square deviation (RMSD) between pairs of transporters (**Figure 1.2E**). Mainly, the structures resolved for the same protein (e.g. LeuT) in different conformations exhibit RMSDs of approximately 2.0 Å in general; those within the same family (e.g. NSS members LeuT, DAT and

MhsT; or BCCT members BetP and CaiT) differ by 3-4.5 Å; while across families (e.g. BCCT and APC family members BetP and AdiC, respectively; or BCCT and NCS1 members BetP and Mhp1) the RMSDs may exceed 6 Å. Thus, although all the transporters have the same fold, there is a hierarchy of structural differences, increasing with their functional differences.

Pairwise alignments of LeuT-fold family sequences confirm their low sequence identities. Pairs belonging to the same family, e.g. DAT-LeuT (NSS), BetP-CaiT (BCCT) or AdiC-ApcT (APC), exhibit sequence identities of 0.25 ± 0.03; across families, the identities drop to 0.15 ± 0.05 (**Figure 1.2F**). If we focus on TM1 and TM6, which are the two most prominent transmembrane domains that undergo significant structural changes during the OF → IF transition, the sequence identities are much higher within families (e.g. 0.60 ± 0.24 for NSS members, and 0.42 ± 0.02 for BCCT members), whereas there is a major drop across families (**Figure 1.2G-H**). For example, CaiT TM1 shows sequence identities of 0.06 ± 0.02 with respect to most transporters. The strong conservation within families and low conservation across families strongly suggest that these helices play a role in defining the specificity of the LeuT-fold transporters.

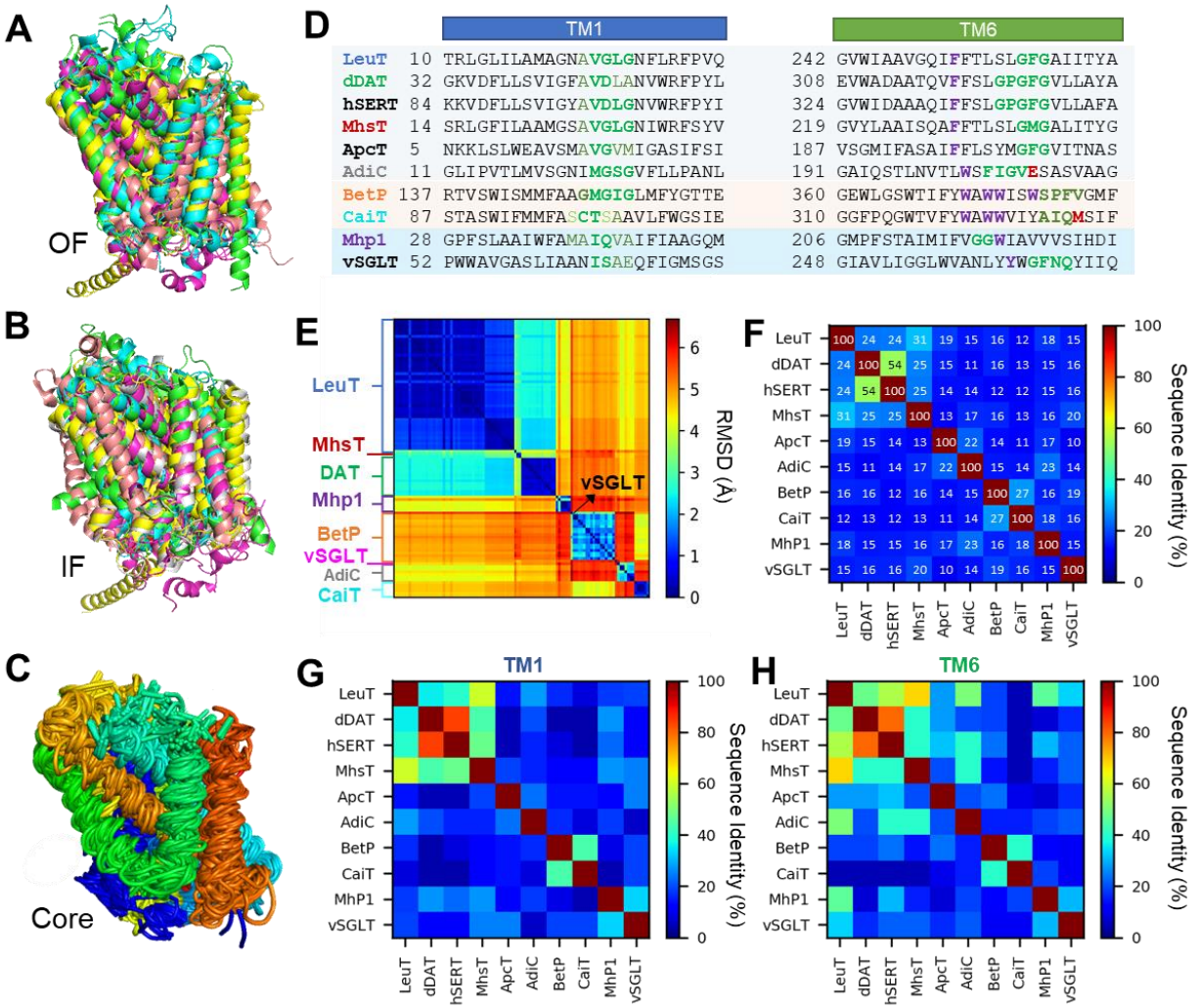

Figure 1.2 Sequence and structural alignment of transporters sharing the LeuT fold. (A-B) Superposition of LeuT-fold transporters in the OF and IF states, colored by transporter type consistent with panel E. **(C)** Superposition of their core region colored by *rainbow colors* from TM1-10 (*blue* to *orange*). **(D)** Sequence comparison of TM1 and TM6 of LeuT superfamily transporters listed in Table **1.1**, plus hSERT and ApcT also sharing the LeuT fold, and belonging to the respective families of NSS and APC. TM1 and TM6 are critical for substrate binding, specificity and translocation. Residues at the helical disruption regions (mostly containing GXG motifs) are highlighted in *green*; and TM6 aromatic residues involved in substrate stabilization are colored *violet*. **(E)** RMSDs between structurally resolved LeuT-fold monomers/protomers. **(F)** Pairwise sequence identities (*color bar* on the right) for all transporters. **(G-H)** Same results if only TM1 or TM6 is considered (figure adapted from (Ponzoni, et al., 2018)).

1.1.2.2 Shared fluctuation profile of core residues: a signature of LeuT-fold dynamics

Previous studies have demonstrated that each protein has its own intrinsic dynamics uniquely encoded by its overall architecture, or fold, which often facilitates its functional interactions; and the intrinsic dynamics may be analytically evaluated using ENMs coupled with NMA (Bahar, et al., 2015; Bahar, et al., 2010). Here, we examined the root-mean-square fluctuation (RMSF) profile of residues derived from the most commonly used elastic network model, the anisotropic network model (ANM) (Atilgan, et al., 2001; Doruker, et al., 2000), for a subset of 11 representative transporters in both OF and IF states, indicated in **Table** 1.1. The results are presented in **Figure 1.3A**. The *black line* therein is obtained for LeuT, rescaled based on X-ray crystallographic B-factors (*purple line*); and the *red line* (and *light red shading*) represents the average behavior (and standard deviation) over the entire set (see **Figure 1.3C** for the behaviors of individual transporters). A strong tendency to exhibit the same "signature" profile among all homologues (monomers and protomers) is seen, with small-to-moderate deviations from the mean.

The next question is to what extent this signature profile is used to enable the global transition of the transporters. To address this question, we determined so-called soft modes, energetically favored by the architecture, which often provide paths for cooperative reorganization of the overall structure and enable allosteric effects. In this case, the availability of structures in both OF and IF states for LeuT, BetP and Mhp1 allowed us to quantitatively assess the structural changes involved in the transition OF →IF (*green line* in **Figure 1.3B**), and compare with ANM soft modes (*red line*). The comparison reveals that motions of the residues during OF→IF transition can be traced back to the global modes, or the signature profile, uniquely defined by the LeuT-fold.

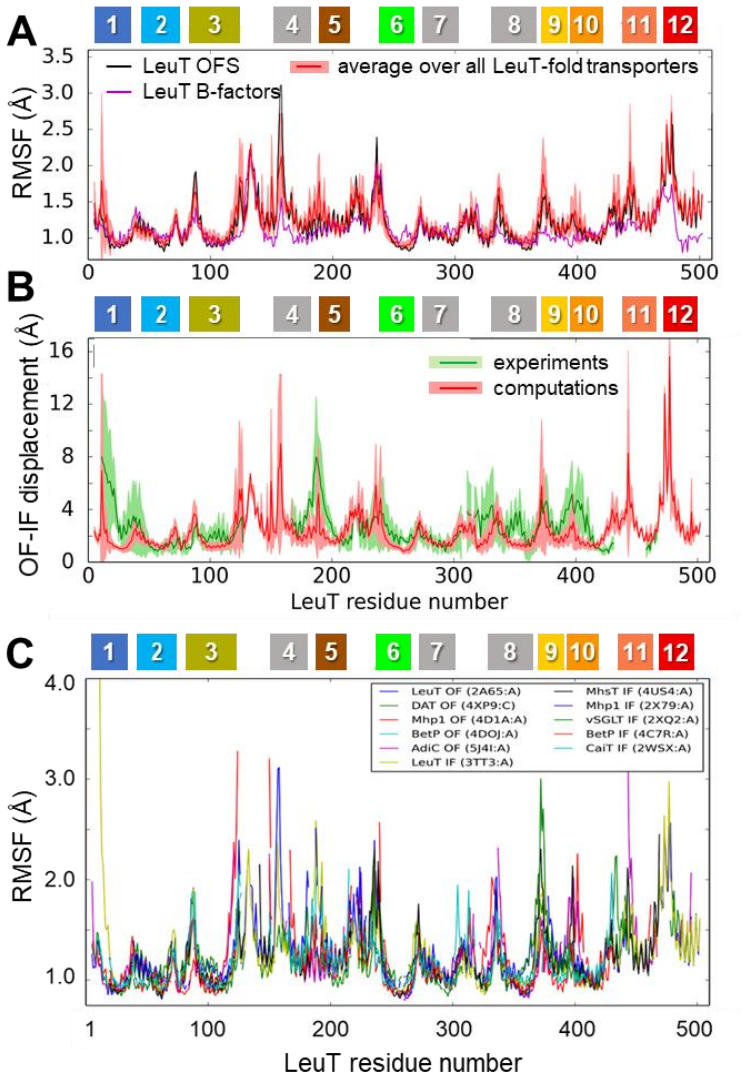

Figure 1.3 Shared dynamics of LeuT-fold residues from theory and experiments. (A) RMSF profile for LeuT (*black curve*) obtained by ANM analysis of OF structure (PDB: 2A65, chain A), is compared to the corresponding B-factor profile from X-ray crystallography (*purple*), and the signature profile (*red curve*) and its standard deviation (*light red band*) computed for a representative set (**Table 1.1**). The correlation coefficient between ANM-predicted RMSFs and those derived from X-ray crystallographic B-factors is 0.65. **(B)** Comparison of experimental (*green*) and ANM-predicted (*red*) global movements (undergone during OF ↔ IF transition) and their variations across LeuT-fold family members. Experimental data refer to LeuT, BetP and Mhp1, resolved in both OF and IF states; ANM profile is obtained from the 10 softest modes evaluated for the representative monomers/protomers. *Shaded areas* indicate the standard deviations. **(C)** ANM-predicted RMSF profiles for 11 structures representing the 8 transporters with LeuT fold in both OF and IF states. This figure is adapted from (Ponzoni, et al., 2018).

A closer inspection shows differences at certain regions, such as TM1, the IC loop between TM4 and TM5, and the extracellular loop EL4 between TM7 and TM8. The latter participates in regulating EC gating and substrate access (Kazmier, et al., 2014), a role involving substrate-specific residues, hence the heterogeneity in the global mode shape at that region. Likewise, the large (approx. 16 Å) displacement of TM1 in the IF state is unique to LeuT (**Figure 1.3C**). This movement is much larger than that observed for TM1 in BetP (Perez, et al., 2012), Mhp1 (Kazmier, et al., 2014) and vSGLT (Watanabe, et al., 2010). Structural comparison shows that BetP TM1a is connected to a long helical segment; but in LeuT, it is connected to a disordered tail and therefore enjoys a high mobility.

1.1.2.3 ANM soft modes characterize conformational variability observed for LeuT superfamily members

Figure 1.3A demonstrates that LeuT-fold monomers or protomers belonging to different functional families exhibit shared dynamics regardless of their conformational (OF/IF) or multimerization (monomer/dimer/trimer) states. Yet, the individual subunits/monomers stabilize the OF, IF or intermediate/occluded state and sample a spectrum of conformational changes, during the transport cycle. How are those different conformers compatible with the same fold and signature fluctuation profile? To gain a mechanistic understanding of the conformational spectrum accessible to LeuT superfamily members, we performed a principal component analysis (PCA) of the ensemble of PDB structures listed in **Table** 1.1. Optimal superposition of 104 monomers and protomers in this set onto the LeuT OF structure (PDB: 2A65; reference structure) permitted us to identify a core region (**Figure 1.1C**) and RMSDs from the mean that varied from ~ 1.5 Å for LeuT monomers/ protomers to ~ 5 Å for vSGLT, BetP and Mhp1. Comparison of the results from PCA with ANM predictions showed that the softest ANM mode (ANM1) computed for the reference

structure (the closest to the average structure of the ensemble in terms of RMSD) yields a cumulative overlap of 0.55 with the principal components 2 and 6 (PC2 and PC6).

Figure 1.4A shows the ensemble of structurally resolved LeuT fold transporters projected onto the theoretically predicted (ANM1) and experimentally supported (PC2 and PC6 combined) principal modes. The correlation is quite high (0.82), confirming that the two sets describe the same direction of deformation. While members of the same family (see the *color code*) tend to cluster together, we note that within each family a certain degree of segregation between IF (*upward triangles*) and OF (*downward triangles*) states takes place, for instance in the case of BetP (in *orange*) and Mhp1 (in *purple*), consistent with the analogous separation for LeuT (*blue*). Such observation points to the fact that a common gating mechanism might be shared among members of the superfamily and is well captured by the softest mode favored by the common fold.

Figure 1.4B-C provides an overview of the "dynamics landscape" of the LeuT-fold transporters. Therein, monomer/protomer structures projected onto the subspace spanned by ANM1, ANM2 and ANM3 allowed visualization of the different classes of proteins based on their collective motions. Notably, proteins belonging to the same functional family tend to cluster, highlighting the relevance of soft modes to transporter function.

Another interesting fact emerges by focusing on the projection of structures onto the first two ANM modes, shown in **Figure 1.4B**. In this representation, a separation can be drawn between trimeric transporters (BetP and CaiT) and monomeric/dimeric transporters, while secondary cuts (*dashed lines*) further subdivide both groups into OF and IF conformations. This separation may be an effect of the structural constraints imposed by the trimeric organization of BetP and CaiT transporters. The trimers feature a different interface compared with dimers, involving the rearrangement of helix H7 in BetP, corresponding to EL3 in LeuT (see *inset* in **Figure 1.4B**). Such

a rearrangement is well reproduced by ANM2 of LeuT (**Figure 1.4B** and **D**) as well as ANM3 (**Figure 1.4E**), suggesting an intrinsic predisposition (via ANM2 and ANM3) of H7 to adopt the correct positioning for trimeric interface formation.

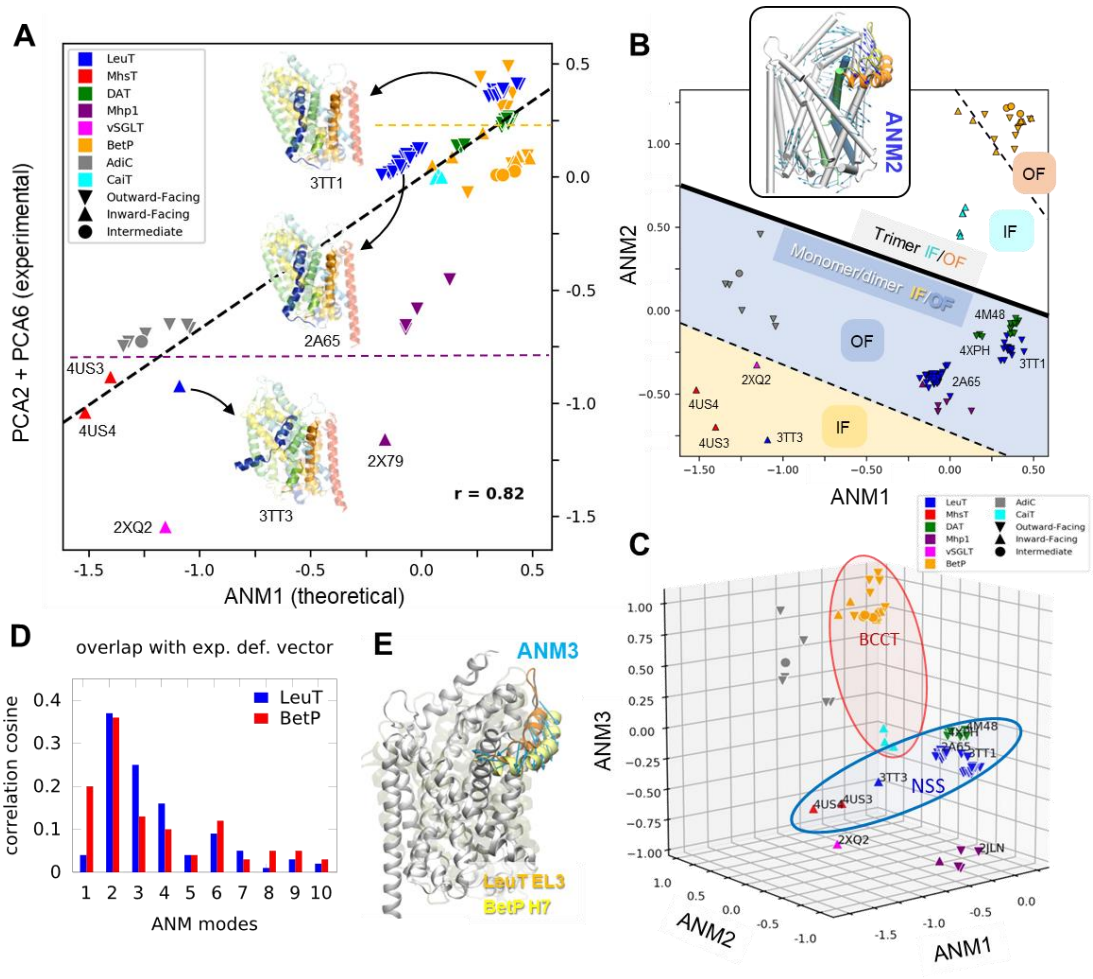

Figure 1.4 Conformational dynamics landscape of LeuT superfamily explains the structural variability experimentally observed for 104 conformers. (A) Projections of the 104 conformers (**Table 1.1**) onto ANM1 and combined mode from PC2 and PC6 yielded a strong correlation (r = 0.82), revealing that the observed differences between these structures comply with the softest mode intrinsically encoded by the LeuT fold. Three LeuT structures from **Figure 1.1B** are displayed to represent the OFo, OFc, and IF states. **(B-C)** Classification of LeuT-fold transporters based on their collective motions. The distribution of the 104 monomers/protomers are displayed in the subspaces of collective modes spanned by two **(B)** and three **(C)** softest ANM modes. ANM2 (see *inset* in **panel B**) directs the reconfiguration of the LeuT EL3 (loop-helix, *yellow*) along a direction (*blue arrows* on the ribbon diagram) in accord with the structural change undergone by the equivalent BetP H7-helix (*orange*) upon trimerization. **(D)** Overlap between the structural variation between the experimental structures of LeuT and BetP, and their respective ANM modes. **(E)** LeuT ANM3 overlaps with the structural difference between LeuT EL3 and BetP H7, as also observed in ANM2 (figure adapted from (Ponzoni, et al., 2018)).

1.1.3 Discussion

The present section focused on a superfamily of structural homologs, LeuT superfamily, that encompasses members from five families of transporters with different functions and low sequence identity. We first characterized their shared structural and dynamic characteristics, and then proceeded to elucidate which features on a local or global scale, structural or dynamic, differentiate them to lead to different functions.

The first task helped us identify a signature residue-fluctuation profile (**Figure 1.3**) intrinsically favored by their common fold, consistent with the postulate that shared fold also implies shared global dynamics. This also implies that those transporters or antiporters have evolved to recruit the same tertiary fold, despite their sequence dissimilarities, presumably driven by the adaptability of the fold to chemical (specificity) and physical (conformational flexibility) differences, thus allowing functional differentiation.

Chemical specificity can be detected at the dissimilar sequence patterns among superfamily members that belong to different functional families, while those transporters within a given family exhibit distinctively higher sequence identities. The difference becomes even more pronounced upon focusing on TM helices involved in substrate/ion binding: similarities among the same functional family members are enhanced, while dissimilarities across different functional families become even more pronounced (**Figure 1.2D-H**). Physical flexibility, on the other hand, is manifested by structural differences between family members on both a global (OF/IF and intermediate states; multiple oligomerization states; **Figure 1.2A-C** and **E**) and a local (helical disruptions, substrate coordination geometry, extrusion of helical loops; **Figure 1.4E**) scale. It is only upon substrate/ligand binding that the pre-existing signature fluctuations are advantageously exploited to drive the transport of substrate. First, EC gate closure is triggered, and then further

insertion of the ion/substrate binding to a structurally irregular, broken helical, region confers a local reordering (e.g. TM1 tilting, or TM6 reorientation) that propagates to the IC region upon the rigidification of the originally frustrated cluster of residues; and the induced structural change opens the IC gate to trigger an influx of IC water, which stimulates the removal of substrate/ion and its release to the IC region. Thus, a cascade of events takes place, stimulated upon substrate and ion binding, typical of the cooperative response of allosteric proteins to ligand binding.

A rigorous examination of the distribution of LeuT superfamily members in the conformational space accessible to them (**Figure 1.4A-C**) demonstrates how the resolved structures are essentially reorganizations of the shared fold along the softest ANM modes 1, 2 and 3 intrinsically favored by their shared fold. ANM1 provides a good description of the principal variations in structure elucidated by the PCA of 104 monomers/protomers; ANM2 plays a dominant role in distinguishing the trimeric transporters; and ANM3 together with ANM2 helps the transition between OF and IF, as evidenced by **Figure 1.4D-E**. This analysis shows that the adaptation of the shared fold to different conformational or oligomerization states is mainly accomplished by the soft paths of reconfiguration intrinsically encoded by the LeuT-fold.

Finally, it is noteworthy to mention that detailed conformational changes that involves the movement of side chains of amino acids usually cannot be captured by coarse-grained ENMs. For example, the unwinding/stretching of TM4-TM5 loop during OF →IF transition of LeuT (Krishnamurthy and Gouaux, 2012), MhsT (Malinauskaite, et al., 2014), Mhp1 (Shimamura, et al., 2010; Weyand, et al., 2008) and BetP (Perez, et al., 2012) cannot be reproduced by the ANM global modes. The unwound part of TM5 in the conserved motif GlyX9Pro of MhsT (Malinauskaite, et al., 2014) and an extension of the TM4–TM5 loop (G258-G263) in hDAT

(Cheng and Bahar, 2015) have been observed to trigger the hydration of Na2, leading to the opening of the IC vestibule.

1.1.4 Methods

1.1.4.1 Preparation of the structural ensemble

We have manually collected 92 PDB structures composed of 104 protomers with the LeuT fold from the PDB. The initial ensemble consists of 50 LeuTs and their mutants, 14 DATs, 2 MhsTs, 6 Mhp1s, 1 vSGLT, 6 BetPs, 4 CaiTs, and 7 AdiCs. 85 protomers were selected after removing the LeuT structures that were almost identical (< 0.3 Å RMSD) to the reference (PDB ID: 2A65) (Yamashita, et al., 2005). See Table 1.1 for more details, and **Figure 1.2** for the respective distributions of pairwise sequence identities, structural RMSDs and biological function among the members of the LeuT family.

For each subfamily (LeuT, DAT, MhsT, Mhp1, vSGLT, BetP, CaiT, AdiC), we first chose the structure with the most complete sequence as family representative to be aligned against the ensemble reference, the one with PDB ID 2A65 (Yamashita, et al., 2005), using the CE structural alignment algorithm. The other PDB structures were aligned to their family representative using sequence alignment algorithms. Through this procedure, we obtained an ensemble consisting of 104 conformers of LeuT fold proteins. We then iteratively superposed all the conformations to minimize their overall pairwise RMSDs.

	Family	Transporter	Conformation [a]	PDB structures [b]
trimer	BCCT	BetP (Koshy and Ziegler, 2015; Perez, et al., 2012; Ressl, et al., 2009)	OF*o*	4LLH
			intermediate	2WIT
			asym (IF*o*/IF*o*/OF*o*)	3P03
			asym (IF*c*/OF/IF*o*)	4AIN
			asym (IF/IF/OF)	**4C7R**
			asym (OF*c*/OF*o*/IF*o*)	**4DOJ**
		CaiT (Kalayil, et al., 2013; Schulze, et al., 2010)	IF*o* (subs. bound)	3HFX
			IF*o*	4M8J, **2WSX**, 2WSW
dimer	APC	AdiC (Gao, et al., 2009; Gao, et al., 2010)	OF*c* (subs. bound)	3L1L
			OF	3NCY, 3LRB, 3LRC, **5J4I**, 5J4N
			intermediate	3OB6
	SSS	vSGLT (Faham, et al., 2008; Watanabe,	IF*po*	**2XQ2**
	NSS	LeuT (Krishnamurthy and Gouaux, 2012; Singh, et al., 2008; Singh, et al., 2007; Yamashita, et al., 2005)	OF*c* (subs. bound)	**2A65**, 2Q6H, 2Q72, 2QB4, 2QEI, 2QJU, 3F3C, 3F3D, 3F3E, 3F48, 3F4I, 3F4J, 3GJC, 3GJD, 3GWU, GWV, GWW, 3MPN, 3MPQ, 3QS5, 3QS6, 3TU0, 3USG, 3USI, 3USJ, 3USK, 3USL, 3USM, 3USO, 3USP, 4HMK, 4HOD
			OF*o*	3TT1, 3F3A, 4MM4, 4MM5, 4MM6, 4MM7, 4MM8, 4MM9, 4MMA, 4MMB, 4MMC, 4MMD, 4MME, 4MMF
			OF	4FXZ, 4FY0, 3QS4
			IF	**3TT3**[c]
monomer	NSS	DAT (Penmatsa, et al., 2013; Wang, et al., 2015)	OF	4M48, 4XNU, 4XNX, 4XP1, 4XP4, 4XP5, 4XP6, **4XP9**, 4XPA, 4XPB, 4XPF, 4XPG, 4XPH, 4XPT
		MhsT (Malinauskaite, et al., 2014)	IF	4US3, **4US4**
	NCS1	Mhp1 (Shimamura, et al., 2010; Weyand, et al., 2008)	OF	2JLN, **4D1A**, 4D1B, 4D1C, 4D1D
			IF	**2X79**

Table 1.1 LeuT fold structures available in the PDB.

[a] OF: outward facing; IF: inward facing; suffixes "*o*" and "*c*" refer to the open or closed states of the EC (in the OF state) or IC (in the IF state) gates; "*po*" is partially open. Structures resolved in the presence of substrate are indicated; [b] PDB codes. Those used in structural alignments (**Figure 1.2A-C**) and in the generation of the fluctuation profiles (**Figure 1.3**), referred to as representative 11 structures (monomers or protomers), are indicated in *boldface*. LeuT may presumably function in either monomeric or dimeric state, but the majority of the resolved structures are dimeric. [c] LeuT IF dimer (Gur, et al., 2015; Zomot, et al., 2015) computed from IF monomer 3TT3. Overall the table contains 104 distinctive conformers belonging to monomers or multimers, which have been used in the present analysis). This table is adapted from (Ponzoni, et al., 2018).

1.1.4.2 Anisotropic network model (ANM)

The Anisotropic Network Model is a broadly used ENM introduced in 2000 (Atilgan, et al., 2001; Doruker, et al., 2000), inspired by the pioneering work of Tirion (Tirion, 1996), succeeding the development of the Gaussian network model (GNM) (Bahar, et al., 1997; Haliloglu, et al., 1997). Given a protein of *n* residues, the minimalist yet powerful setup for the ANM is to represent each residue by a node whose initial position is identified by the coordinate of the α-carbon of the residue. Then a spring is connected to two nodes if the distance between the α-carbons of the corresponding residues is smaller than a cutoff d_0. Each spring represents a harmonic potential whose equilibrium position is the initial position of the two connected nodes, i.e. coordinates of the α-carbons of the residues in the native conformation. The overall ANM potential of the network is defined as

$$V_{ANM} = \frac{1}{2}\sum_{i,j} \gamma\left(d_{ij} - d_{ij}^0\right)^2, \tag{1.1}$$

where d_{ij} and d_{ij}^0 are the instantaneous and equilibrium (scalar) distances, respectively, between nodes i and j. In the minimalist setup, a uniform force constant γ is used for all the springs. However, this can be easily modified for more sophisticated models (an example of which we will see in Section 2.0). In this study, we built the ANMs for 104 LeuT fold conformers which are aligned and trimmed or extended to the same number of residues (gaps are filled by dummy atoms whose coordinates are represented by the average coordinate of the other family members). This allows the ANM results to be directly comparable at the cost of introducing minimal computational artifacts.

Next, the ANM can be solved efficiently by eigenvalue decomposition of a $3n \times 3n$ Hessian matrix, **H**, the off diagonal super elements (3×3 submatrices) of which are the second derivatives

of the ANM potential (equation 1.1) with respect to three degrees of freedom of every interaction pair of nodes evaluated at their equilibrium coordinates (x_{ij}^0, y_{ij}^0, z_{ij}^0; see (Atilgan, et al., 2001) for the derivation),

$$\mathbf{H}_{ij} = -\frac{\gamma}{d_{ij}^0}\begin{bmatrix} \left(x_{ij}^0\right)^2 & x_{ij}^0 y_{ij}^0 & x_{ij}^0 z_{ij}^0 \\ x_{ij}^0 y_{ij}^0 & \left(y_{ij}^0\right)^2 & y_{ij}^0 z_{ij}^0 \\ x_{ij}^0 z_{ij}^0 & y_{ij}^0 z_{ij}^0 & \left(z_{ij}^0\right)^2 \end{bmatrix}, \tag{1.2}$$

and the diagonal super elements are evaluated as the negative sum the off-diagonal super elements in the same row (or column). Eigenvalue decomposition of **H** yields 3*n* - 6 non-zero eigenvalues ($\boldsymbol{v}_k$) and eigenvectors (λ_k), if the network is well connected (no disjoint components). The eigenvectors are the *normal modes* which form an orthogonal basis set spanning the motion space, and each describes the displacements of the network nodes along that particular mode. The eigenvalues are proportional to the square of the mode frequencies and represent the curvature of the harmonic energy function along the mode "axis". Therefore, displacements of a given size along high frequency (HF) modes (*fast modes*) are energetically more expensive than those along the low frequency (LF) ones (*slow modes*). According to the equipartition theorem, the total vibrational energy is on average equally distributed on each mode, so the network nodes experience greater displacements along the slow modes than the fast. Because of this, slow modes are of greater interest and importance for analyzing and characterizing the equilibrium dynamics of the molecule.

A 3*n* × 3*n covariance matrix* that can be evaluated by taking the pseudoinverse of **H**:

$$\mathbf{C}_{ANM} \sim \mathbf{H}^{-1} = \sum_{i=1}^{k} \frac{1}{\lambda_i} \boldsymbol{v}_i \boldsymbol{v}_i^T , \tag{1.3}$$

where $1/\lambda_k$ is the variance (or vibrational amplitude) of the mode *k*. Typically, *k* = 3*n* – 6 for summing over all the modes. A lower value of *k* can be used to obtain a low-rank approximation

of the covariance matrix for computational efficiency and elimination of noise. A good choice of *k* can be determined by examining, for example, if the *cumulative weight* of the first *k* modes is over ~80%,

$$fvar(k) = \frac{\sum_{i=1}^{k} \frac{1}{\lambda_i}}{\sum_{j=1}^{n-6} \frac{1}{\lambda_j}}. \tag{1.4}$$

The elements in the covariance matrix can be viewed as $n \times n$ blocks of 3×3 submatrices (similar to those of **H**), and each submatrix $\mathbf{C}_{ij}$ corresponds to the covariances between the movements of node *i* and *j* along the three degrees of freedom (every combination of *x*, *y*, *z* and therefore each submatrix is 3×3). In many applications, it is more useful to examine the *cross-correlations* $\tilde{\mathbf{C}}$ between the vectorial (3D) displacements of the two nodes. In contrast to the covariance matrix, each element $\tilde{\mathrm{C}}_{ij}$ of thus $n \times n$ cross-correlation matrix $\tilde{\mathbf{C}}$ is a scalar representing the cross-correlation $\langle \Delta \boldsymbol{r}_i \cdot \Delta \boldsymbol{r}_j \rangle$ between the vectorial displacements of nodes *i* and *j*. This allows one to quickly identify dynamically coupled node pairs, and the diagonal elements represent the *mean-square fluctuations* (MSFs) of the network nodes, $\langle (\Delta \boldsymbol{r}_i)^2 \rangle$. The MSFs, as well as the covariance matrix are inversely proportional to the force constant γ, and proportional to the absolute temperature *T*. The precise value of γ is usually unknown, and adjusted to achieve quantitative agreement with experimental measurements such as the X-ray crystallographic B-factors. However, it is usually unnecessary to obtain the absolute scale in practice because, the distribution of MSFs among the residues is invariant to γ, as well as the *cross-correlations* between the displacements,

$$\tilde{\mathbf{C}}_{ij} = \frac{\langle \Delta \boldsymbol{r}_i \cdot \Delta \boldsymbol{r}_j \rangle}{\langle (\Delta \boldsymbol{r}_i)^2 \rangle \langle \left(\Delta \boldsymbol{r}_j \right)^2 \rangle} \tag{1.5}$$

For example, in the present comparative analysis of protein family members' dynamics, the coarse-graining level allows us to safely assume that all members have comparable volumes, packing densities, and physiological conditions, and thus we can normalize the MSFs to observe the so-called *mobility profiles* of residues (i.e. normalized distributions of MSFs).

1.1.4.3 Principal component analysis (PCA) of protein conformations

PCA has been widely used in various disciplines. In the application to molecular simulations, PCA has been broadly used for extracting the so-called essential dynamics (Amadei, et al., 1993) of proteins, using conformations (also called snapshots or frames) generated in molecular dynamics (MD) trajectories. Other applications of PCA to explore the conformational space of proteins include the application to structural models observed in nuclear magnetic resonance (NMR) experiments, or to ensembles of PDB structures of the same protein in different states, etc. Essential dynamics refers to the first few dominant principal components (PCs) which are often found to be the essential modes of motions associated with molecular functions.

Upon superposing *m* different structures (conformers) to a reference, a covariance matrix of residue fluctuations can be constructed by

$$\mathbf{C}_{PCA} = \frac{1}{m}\sum_{i}(\boldsymbol{q}_i - \langle \boldsymbol{q} \rangle)(\boldsymbol{q}_i - \langle \boldsymbol{q} \rangle)^T , \tag{1.6}$$

where $\boldsymbol{q}_i$ is the generalized coordinate of the residues in the i^{th} state. $\langle \boldsymbol{q} \rangle = \frac{1}{m}\sum_i \boldsymbol{q}_i$. Similar to the ANM analysis, the eigenvalue decomposition of $\mathbf{C}_{PCA}$ yields a set of modes of motions with associated variances. The modes are shown to be directly comparable to those obtained by the ANM (Bakan and Bahar, 2009); see Section 1.1.2.3. However, it should be noted that PCA only yields $m - 1$ meaningful modes (non-zero eigenvalues), as opposed to $3n - 6$ in the ANM, if the number m of states (samples) is smaller than $3n$, so that the motion space spanned by the PCs may

be incomplete, depending on the input data. Nonetheless, PCA is a powerful tool for extracting and identifying functional modes of motions and for quantitative comparison of experimental and theoretical (ANM) results.

1.1.4.4 Projection of conformations onto subspaces spanned by the ANM/PCA modes

Given the conformational change of a structure with respect to the reference $\Delta\boldsymbol{q}$ (generalized coordinate), and a set of mode vectors $V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_k]$ obtained for the same structure from either the ANM or PCA, the projection p_i of the conformational change $\Delta\boldsymbol{q}$ onto the mode *i* can be calculated by their inner product:

$$p_i = \Delta\boldsymbol{q}^T\boldsymbol{v}_i. \tag{1.7}$$

Note that $\boldsymbol{v}_i$ is a unit vector so in the extreme case of $\Delta\boldsymbol{q}$ perfectly aligned along $\boldsymbol{v}_i$, $p = |\Delta\boldsymbol{q}|$. Finally, $(p_1, p_2, \dots, p_k)$ forms a new coordinate for projecting the conformational change $\Delta\boldsymbol{q}$ onto the subspace spanned by the mode set *V* (see **Figure 1.4A-C** for illustration).

1.1.5 Acknowledgment

The content of this subsection was published (Ponzoni, et al., 2018) with me being one of the first three coauthors. The other two are Drs. Luca Ponzoni and Mary Hongying Cheng who contributed significantly to this work. I performed all the computations and assisted in data analyses and interpretation. Dr. Ivet Bahar conceived the presented idea and supervised the project.

1.2 Signature Dynamics of Protein Families

1.2.1 Introduction

In recent years, there has been an increasing interest in interpreting sequence evolutionary trends in the light of biophysical models, reconciling evolutionary biology and structural biophysics (Echave and Wilke, 2017; Liberles, et al., 2012). In Section 0, we demonstrated how the ENMs may be used to analyze and reveal the shared dynamics of the sequentially divergent LeuT fold proteins. The efficiency and versatility of the ENMs allow them to be generalized and extended to study larger proteins and more populated families of proteins, which may bring insights into the evolution of protein structure and dynamics on a broader scope.

We recently introduced a new interface, SignDy (http://prody.csb.pitt.edu/signdy/), for evaluating the Signature Dynamics of protein (super)families (Zhang, et al., 2019). The hypothesis was that similar to signature motifs at the sequence and structure levels, each family might be characterized by a signature dynamics. As will be shown below, application to 116 superfamilies disclosed basic principles for functional fitness and diversification: exploiting the robust global dynamics of a versatile fold, and gaining specificity via localized, yet impactful, fluctuations conserved among subfamily members but divergent across subfamilies.

We further illustrated the utility of SignDy by way of application to three families of folds: 1) leucine transporter (LeuT), 2) periplasmic-binding protein type-1 (PBP-1), and 3) triosephosphate isomerase (TIM) barrel (see **Figure 1.5** for an overview of the ensemble composition and their sequence and structure similarities). SignDy proved to be an effective tool for quantitative evaluation of both generic dynamics of families, and specific dynamics of subfamilies, identifying the specific modes of motions that distinguish subfamilies (shared by

subfamily members but sharply different across subfamilies), and learning how evolutionarily selected folds exploit collective modes of motions in different frequency regimes to reconcile a diversity of sequences and functions with the same architecture. The results reveal the conservation/differentiation of structural dynamics in relation to the evolution of sequence and structure.

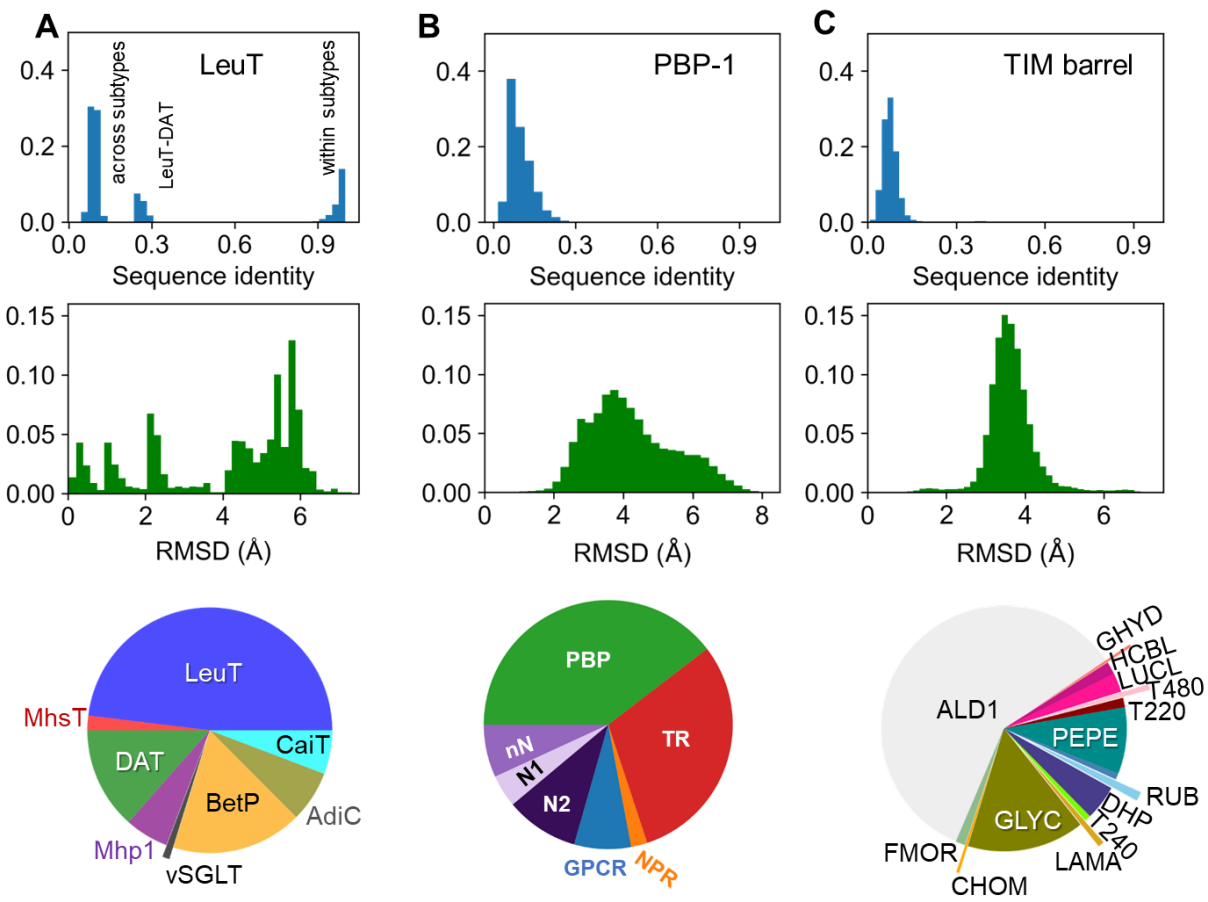

Figure 1.5 Sequence, structure and function properties of LeuT, PBP-1 and TIM barrel fold family members. Distributions of average fractional sequence identities and average structural RMSDs within the **(A)** LeuT, **(B)** PBP-1, and **(C)** TIM-barrel fold families (*upper* and *middle* panels), and biological functions of family members (pie charts, *lower* panels). The sequence identity histogram for the LeuT family of transporters (A) shows three groups, at 0.10, 0.28 and 0.95, which correspond to the low similarity between distinct subtypes, the intermediate similarity between LeuT/DAT pairs, and the near identity of transporters of the same subtype. PBP-I and TIM barrel family members (B-C) have highly dissimilar sequences (with average sequence identity of 0.11 and 0.085, respectively), while their average RMSD values are 4.25 and 3.64 Å. Abbreviations: GLYC, glycosidases; AdiC, arginine/agmatine antiporter; ALD1, Aldolase class I; BetP, glycine betaine transporter; CaiT, carnitine/butyrobetaine antiporter; CHOM, copper homeostasis (CutC) domain; DAT, dopamine transporter; DHP, dihydropteroate synthase-like; FMOR, FMN-linked oxido-reductase; GHYD, glycoside hydrolase, family 3, N-terminal domain; GPCR, G-protein coupled receptor; HCBL, homocysteine-binding-like domain; LAMA, D-lysine 5,6-aminomutase a-subunit; LeuT, leucine transporter;

LUCL, luciferase-like domain; Mhp1, benzyl-hydantoin transporter; MhsT, multi-hydrophobic amino acid transporter; MHYD, metal-dependent hydrolases; MTMB, monomethylamine methyltransferase MtmB; nN, non-NMDA iGluR; N1, GluN1 NMDA iGluR subunit; N2, GluN2 NMDA iGluR subunit; PBP, periplasmic binding protein; PEPE, phosphoenolpyruvate-binding domain; RUB, ribulose bisphosphate carboxylase large subunit C-terminal domain; T220, TIM barrel superfamily 3.20.20.220; T240, TIM barrel superfamily 3.20.20.240; T480, TIM superfamily 3.20.20.480; T540, TIM barrel superfamily 3.20.20.540; TR, transcription regulator; vSGLT, Vibrio sodium/galactose transporter. This figure is adapted from (Zhang, et al., 2019).

1.2.2 Results

As a proof of concept, we first explored and tested the utility of SignDy by way of application to three families of folds: 1) LeuT, 2) periplasmic-binding protein type-1 (PBP-1), and 3) triosephosphate isomerase (TIM) barrel. Then, we proceeded to perform a systematic computational analysis of 26,899 proteins belonging to 116 CATH superfamilies.

1.2.2.1 Transport and multimerization mechanisms of LeuT fold proteins favored by their signature dynamics

In addition to the results obtained in Section 0, here, we focus on transport and multimerization mechanisms of LeuT members. **Figure 1.6** reveals how the three global modes operate in a complementary way to enable substrate transport: they divide the fold into two parts from three orthogonal perspectives, resulting each in concerted opposite-direction (anticorrelated) fluctuations (or breathing motions) of the respective parts. Their combination allows for the cooperative opening and closing of the central substrate/ion-binding pocket (**Figure 1.7B**). The close-to-zero values in **Figure 1.6A** (indicated by *vertical pink shades*) indicate pivotal sites at the interface between oppositely moving substructures. Notably, the first pivotal site (LeuT residue

number 21-26, DAT 44-48, BetP 149-154, etc. See **Figure 1.2D** for a sequence alignment) is the conserved broken region of TM1 helices among LeuT-fold proteins. They are characterized in the previous study (Ponzoni, et al., 2018) and found to harbor critical binding sites for the substrates and ions (**Figure 1.7**).

Closer examination reveals large displacements in EC loop 3 (EL3; known as helix H7 in BetP and CaiT) (*black arrows* in **Figure 1.6A** and **C**). The transporters exhibit large structural heterogeneities at this region (**Figure 1.8**). However, the movement of EL3 is not random. On the contrary, it is driven by a cooperative mode (ANM2) that enables the transition between OF and IF states of the transporter; and further motion of BetP H7 along the same direction/mode allows for inter-subunit contacts that stabilize the trimer (**Figure 1.8C**; also see Section 1.1.2.3 and **Figure 1.4**). This is further supported by the observation that H7 suppressed in most IF conformers except in vSGLT, BetP, and to some extent CaiT, where this specific region facilitates trimerization (**Figure 1.8D**).

Another region distinguished by its conformational adaptability is IC loop 2 (IL2; *red arrow* in **Figure 1.6A** and **C**). This region undergoes large rearrangements during the OF ↔ IF transitions of LeuT (Krishnamurthy and Gouaux, 2012), Mhp1 (Shimamura, et al., 2010), MhsT (Malinauskaite, et al., 2014), and BetP (Perez, et al., 2012), the directions and the sizes of the deformations varying between members. The fluctuations are prominent in the IF states of LeuT, Mhp1 and CaiT, but not in the IF state of MhsTs, BetP and vSGLT nor the OF states. The departure from the generic signature profile at this region suggests a role in imparting specificity (**Figure 1.6A** and **C**).

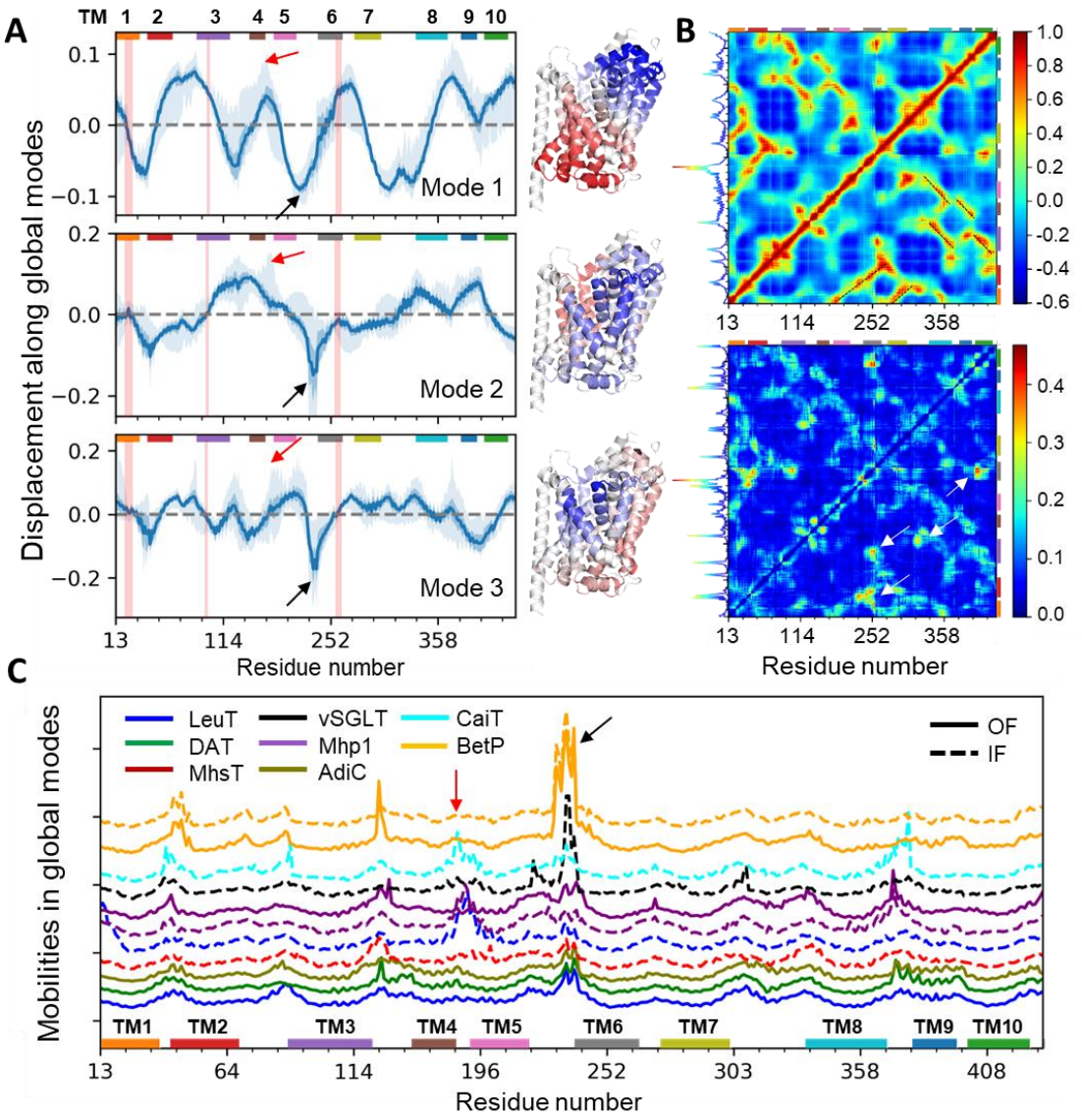

Figure 1.6 Generic and specific features of LeuT fold dynamics. Displacements along the global modes shared by family members (mean profiles, *solid curves*), their differentiation (standard deviations; *darker shaded area*), and the full range of variations (*lighter shaded area*). The *ribbon diagrams* generated for a representative LeuT structure (PDB ID: 2A65) are color-coded (from *blue* to *red*) by the size and direction of motions (from negative to positive) in each mode. **(B)** Generic covariance map (*top*) and its standard deviation (*bottom*), based on $k \leq 20$ modes. Specific residue pairs whose cross-correlations significantly depart from the average are indicated by *white arrows* (*bottom*). The *curve* along the *left ordinate* shows the row-average. **(C)** Detailed view of the global/soft motions ($k \leq 5$) for 13 representative structures from 8 transporter families (*labeled*), in IF (*dashed*) and/or OF (*solid*) states. The *curves* are vertically shifted for visual clarity. This figure is adapted from (Zhang, et al., 2019).

The cross-correlations maps (**Figure 1.6B**) highlight the structural elements that undergo coupled same-sense (*red*) or opposite-sense (or anticorrelated; *blue*) motions. The largest variations in cross-correlations (*lower* map in **panel B**) take place in the motions of TM6 with respect to TMs 1-3 and 10, suggesting a driving role in eliciting cooperative changes. These interhelical distances have been noted to define the extent of opening/closure of the EC and IC vestibules (Cheng and Bahar, 2014; Drew and Boudker, 2016). TM1 movements are shown here to be anticorrelated with respect to TM10 which forms a coherent block with TM5 and TM7. These observations are consistent with recent H/D exchange mass spectrometry experiments where partial unwinding of TM1, 5, 6 and 7 drives the OF → IF transition (Merkle, et al., 2018).

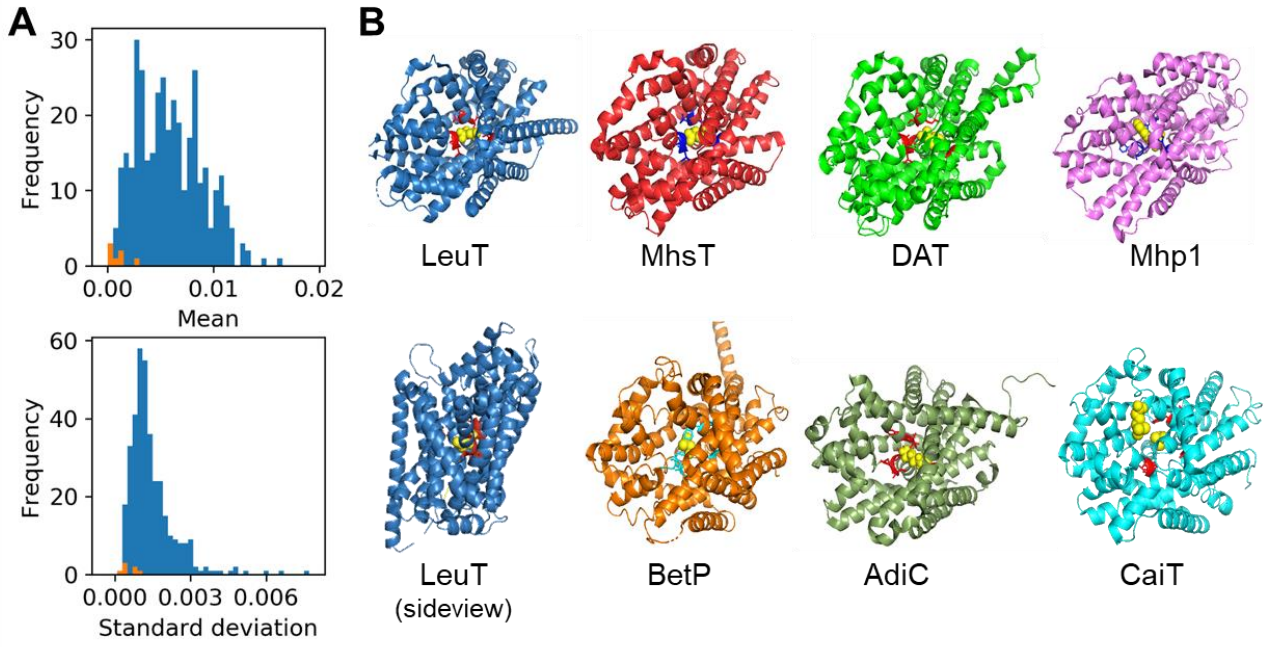

Figure 1.7 The substrate-binding pocket of LeuT-fold transporters shows minimal fluctuations. (A) Distribution of the mean values (*top*) and standard deviations (*bottom*) for square fluctuations of all residues (*blue bars*) and only substrate-binding site residues (*orange bars*). **(B)** Structures of most transporter subtypes are shown from the extracellular side to illustrate the binding pocket with the presence of the substrate (if available). The corresponding PDB codes are: 2A65 for LeuT, 4US4 for MhsT, 4XP9 for DAT, 4D1D for Mhp1, 2XQ2 for vSGLT, 4LLH for BetP, 5J4I for AdiC, and 4M8J for CaiT.

Using sequence-, structure- and dynamics-based distance metrics, we generated the maps and dendrograms presented in **Figure 1.9** for the LeuT family. While the topologies of the three trees are similar, an increased discrimination between family members can be seen as one proceeds from sequence, to structure, to dynamics, along with the re-distribution of selected members, highlighted in ellipses, based on dynamics. **Panel A**, based on sequence, obviously collapses all leucine transporters into a single node regardless of their conformational state; panel **B** separates the OF and IF conformers of LeuT, clustering the latter with two MhsT structures that are also in the IF state. In panel **C**, on the other hand, IF LeuT conformers are clustered with IF Mhp1, CaiTs and vSGLT, presumably sharing similar dynamics, while MhsT is differentiated despite its structural similarity. Such discriminative power can be explained by the projection of LeuT family members onto the ANM signature modes 1-3 (**Figure 1.9C**).

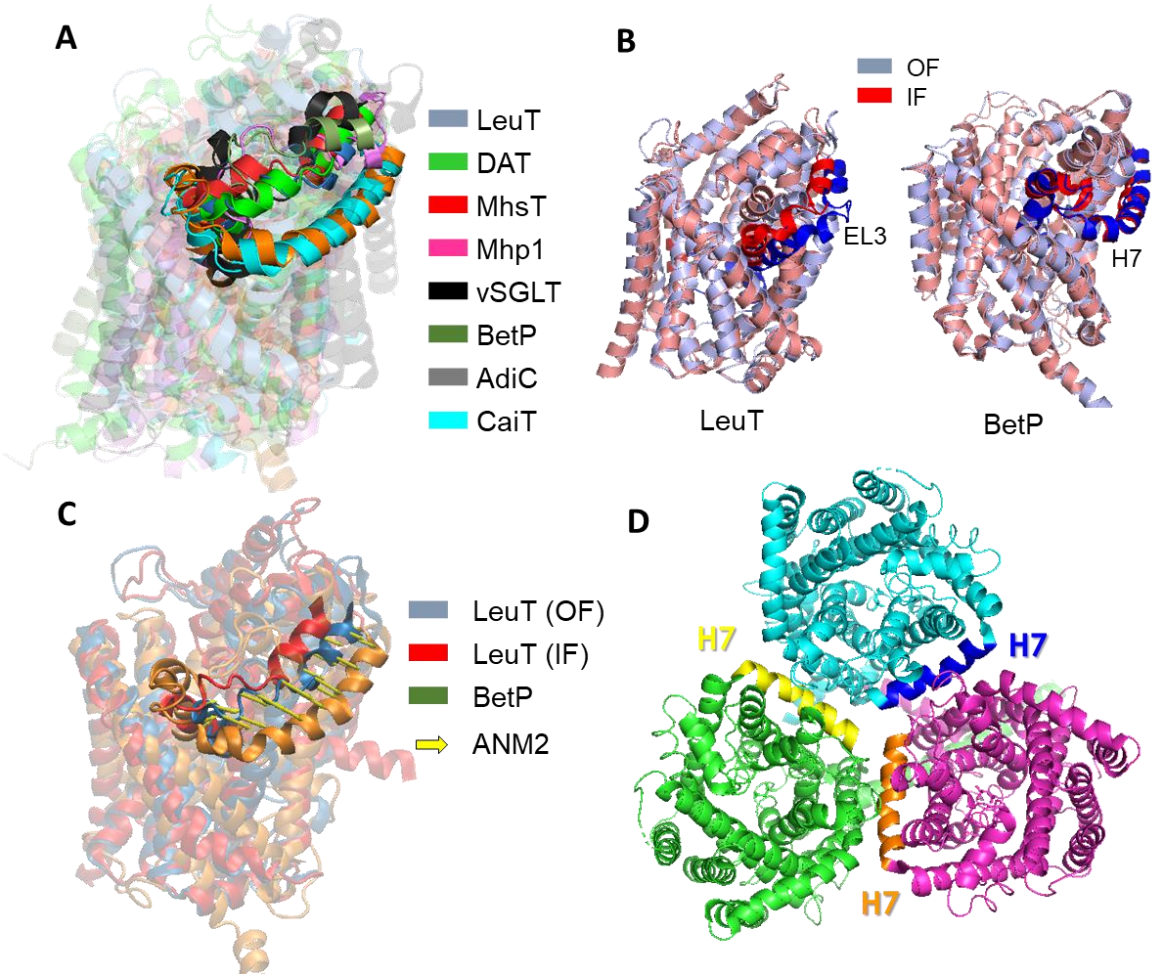

Figure 1.8 Structural differences of LeuT EL3/BetP H7 explained by the ANM and multimerization state. (A) Superposition of representative structures from each family. The region of LeuT EL3 or equivalent BetP H7 is highlighted. BetP and CaiT (both trimeric) are distinguished from other members of the LeuT fold family. **(B)** Comparison of the OF and IF structures of LeuT and BetP, respectively. **(C)** Alignment of LeuT OF and IF structures and a BetP structure. The structural change in this region, predicted by ANM mode 2 calculated on a single LeuT OF structure (PDB ID: 2A65), is indicated by the *yellow arrows*. **(D)** BetP trimer with each protomer colored differently and the H7 helices highlighted by a different color. This figure is adapted from (Zhang, et al., 2019).

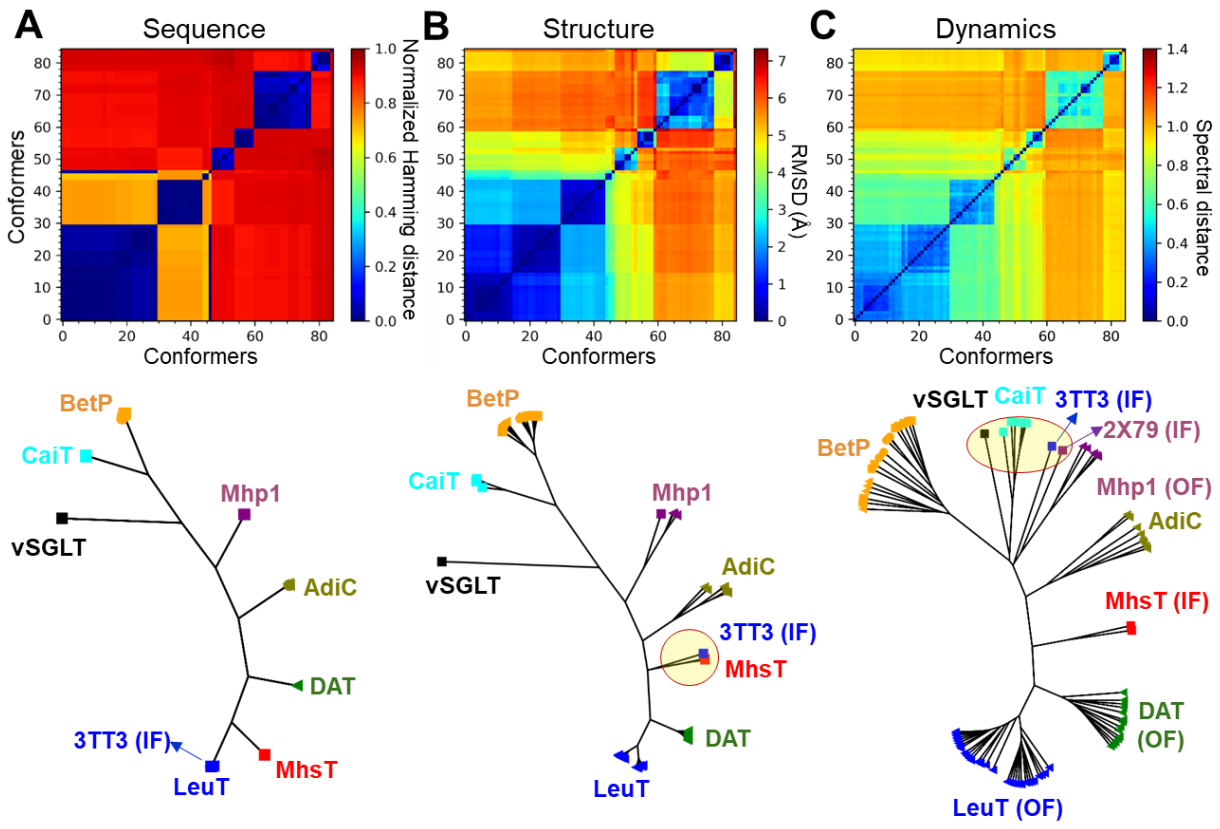

Figure 1.9 Clustering of LeuT superfamily members based on their variations in sequence, structure and dynamics. Distance matrices (*upper*) and corresponding UPGMA trees (*lower*) of LeuT folds based on **(A)** sequence distance, **(B)** RMSD and **(C)** spectral distance calculated based on top 5 GNM modes obtained for each conformation. In the matrices, columns or rows represent conformers in the ensemble clustered and ordered based on pairwise RMSDs shown in panel B (see **Table 1.1** for the order of conformers). In the trees, each node represents a conformer in one of the three states: OF (*triangle*), IF (*square*) or intermediate (*circle*). Note that the nodes in sequence and structure trees are overlapping which makes them look less populated. The node for a LeuT IF structure (PDB ID: 3TT3) is highlighted in the tree diagrams, as its locations are different in these trees. The node for the only IF structure of Mhp1 (PDB ID: Mhp1) is highlighted in the dynamical tree. The trees are generated using the software interactive Tree Of Life (iTOL) (Letunic and Bork, 2007; Letunic and Bork, 2019). The figure is adapted from (Zhang, et al., 2019).

1.2.2.2 Signature dynamics illustrated for three protein superfamilies

Figure 1.10A-C illustrates the signature dynamics for three folds, LeuT, PBP-1 and TIM barrel. Information on the corresponding datasets of proteins can be found in the respective Supplementary Tables S1-3 in (Zhang, et al., 2019); their sequence, structure and function distributions are presented in **Figure 1.5**. The average mobility profile of residues resulting from global modes of motion (up to $k = 3$ (*blue*)) and LF motions (up to $k = 10$ (*orange*) and 20 (*green*) modes) are displayed, along with their standard deviation within each family. Minima and maxima can be traced back to secondary structural elements (indicated by *colored bars* along the *abscissa* in **Panel A** and **C**) and loops (or disordered regions, respectively. This is due to the high packing density at secondary structural elements manifested by small-amplitude fluctuations at those regions. The minimal difference between the three curves in each panel indicates the robustness of the signature dynamics defined by global modes. The LF modes in the range $10 \leq k \leq 20$, which are usually less collective than those in $k \leq 10$, induce increased variations (*shades*) indicative of a differentiation among members while preserving the signature dynamics.

To assess the level of conservation of global modes within families, we evaluated the mode-mode correlation cosines averaged over all family members, $\langle cc_k \rangle$, for each equivalent mode k. The results are presented in **Figure 1.11 A-C** (*green curves* and *shades* for the respective averages and standard deviations). Sharp peaks at the lowest frequency end of the spectra and rapid decays with increasing mode number confirm the conservation of global modes.

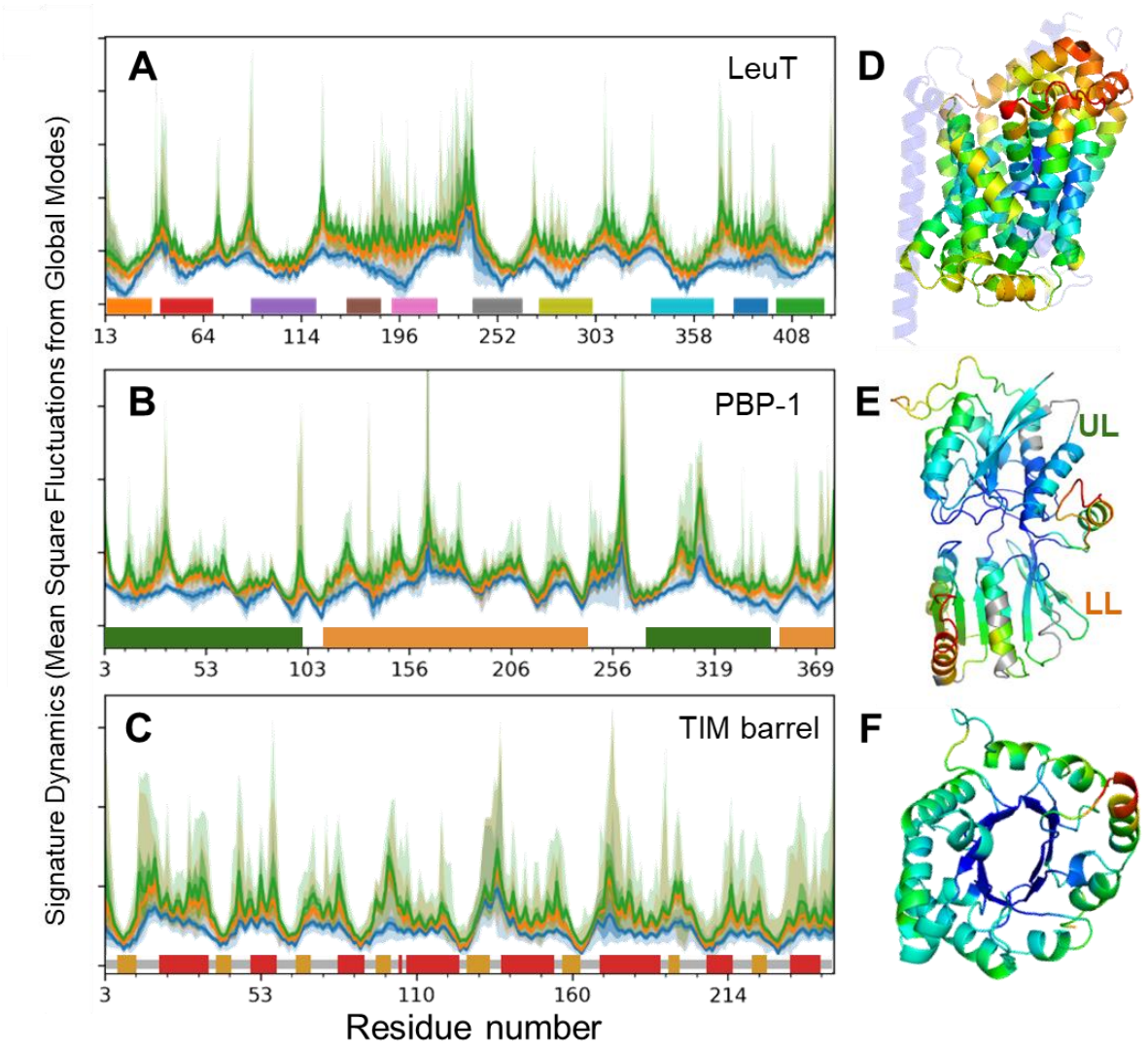

Figure 1.10 Signature-dynamics of each family is robustly defined by global motions uniquely defined by the fold.(A-C) Results for the respective fold families LeuT, PBP-1 and TIM barrel. Mobility profiles driven by k = 3 (*blue*), 10 (*orange*) and 20 (*green*) modes are presented, along with their standard deviations (bands in lighter shades). Horizontal bars along the abscissa indicate **(A)** the transmembrane (TM) helices of LeuT, **(B)** the upper lobe (UL) and lower lobe (LL) of PBP-1, and **(C)** the secondary structure (*orange*, α-strands; *red*, β-helices) of TIM barrel. **(D-F)** Ribbon diagrams of representative members, with core residues color-coded by their mobilities in global modes ($1 \leq k \leq 3$; *blue*, minimal; *red*, maximal). This figure is adapted from (Zhang, et al., 2019).

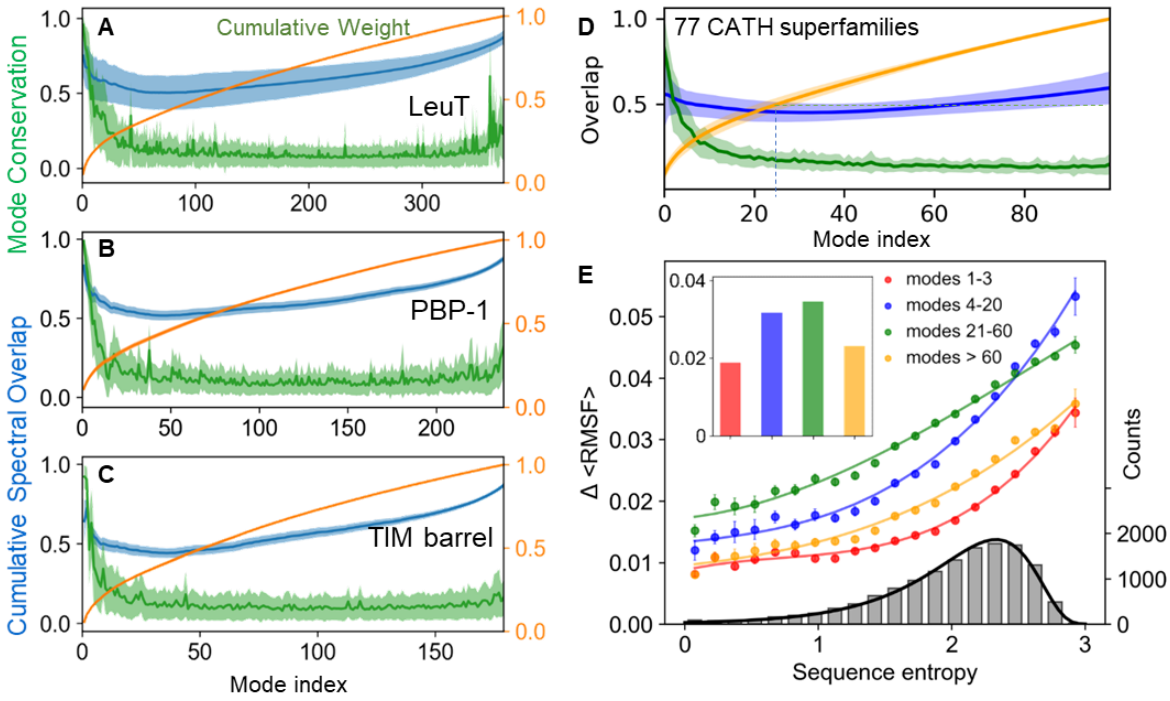

Figure 1.11 Mode conservation and spectral overlap analysis shows the high conservation of global modes and differentiation of LTIF modes among family members. (A-C) Mode conservation profile given by mode-mode correlation cosines $\langle cc_k \rangle$ averaged over all family members (*green*), cumulative spectral overlaps (*blue*), and cumulative weights of individual modes (*orange*) plotted as a function of mode index for LeuT, PBP-1 and TIM-barrel folds, respectively. The *curves* display the averages over all members in each family and the *bands* show the standard deviations. In all three cases, the mode conservation decreases sharply from 0.96 ± 0.03 for mode 1, to 0.63 ± 0.23 for mode 5, and 0.18 ± 0.15 for mode 30. **(D)** Same result for first 100 modes obtained for 77 CATH superfamilies with number of residues greater than 100. The range $1 \leq k \leq 100$ covers four regimes of motions: global/softest ($k \leq 3$), low frequency (LF) ($4 \leq k \leq 20$), LTIF ($21 \leq k \leq 60$) and high frequency (HF) ($k \geq 60$). **(E)** Change in root-mean-square fluctuations, ΔRMSF, computed for all residues in each of the 77 CATH superfamilies as a function of sequence entropy evaluated for four frequency regimes (labeled). The corresponding average values are shown by colored bars in the *inset*. The *colored curves* are weighted least square fits to computed data using cubic regression, with respective correlation coefficients >0.99. The distribution of sequence entropy for the 77 superfamilies, shown by the *gray bars* (*right ordinate*) with a bin size of 0.15 and an average value is 2.0, fits a lognormal probability distribution (*black curve*) with a correlation coefficient of 0.997. This figure is adapted from (Zhang, et al., 2019).

1.2.2.3 Robust global modes define signature dynamics

To confirm the dominance of global modes as a determinant of family signature-dynamics, we examined their level of conservation within CATH superfamilies. To this aim, we considered 116 highly populated superfamilies which overall include 26,899 PDB structures (data not shown; see Supplementary Table S4 in (Zhang, et al., 2019)). For each superfamily, we calculated the pairwise sequence identities, RMSDs, and spectral distances between all pairs of members and evaluated the average values and standard deviations. The histograms in **Figure 1.12** show the sequence, structure, and spectral (dis)similarities among the members of each superfamily. The average of pairwise sequence identities within the superfamilies are around 0.2 (or an average of sequence distances of ~0.8, equivalently) meaning that the sequences within superfamilies are quite divergent; whereas the average RMSDs are ~4.0 Å, indicating strong structural homology within superfamilies (especially since we have already filtered out similar structures, see Section 1.2.4.1). Nonetheless, we found a correlation between sequence identity and RMSD (r = 0.61; **Figure 1.12A**), suggesting the fact that divergent sequences still tend to imply divergent structures, to a degree.

Next, for each superfamily, we computed the mode-mode correlation cosine curves, and then evaluated the average over all superfamilies. The resulting master curve and its standard deviation (shown in **Figure 1.11D**, *green curve* and *shade* for $1 \leq k < 100$) consistently show that global modes are highly conserved. The average correlation cosine for the top-ranking mode (k = 1) of superfamily members is 0.80 ± 0.19 and drops to 0.20 ± 0.07 for k = 20. Higher modes display a plateau with minimal (0.1 - 0.2) correlation.

Larger proteins/domains have access to a broader conformational space and a wider spectrum of motions. One might expect their dynamics to be more heterogeneous, leading to

weaker mode conservation among members. Computations showed, however, that the dependency of mode conservation propensity on protein size is minimal (data not shown; see Supplementary Figure S3 in (Zhang, et al., 2019) instead). The top-ranking modes exhibit strong correlations, irrespective of the size of the protein, again confirming that a handful of global modes robustly define the signature dynamics of the family.

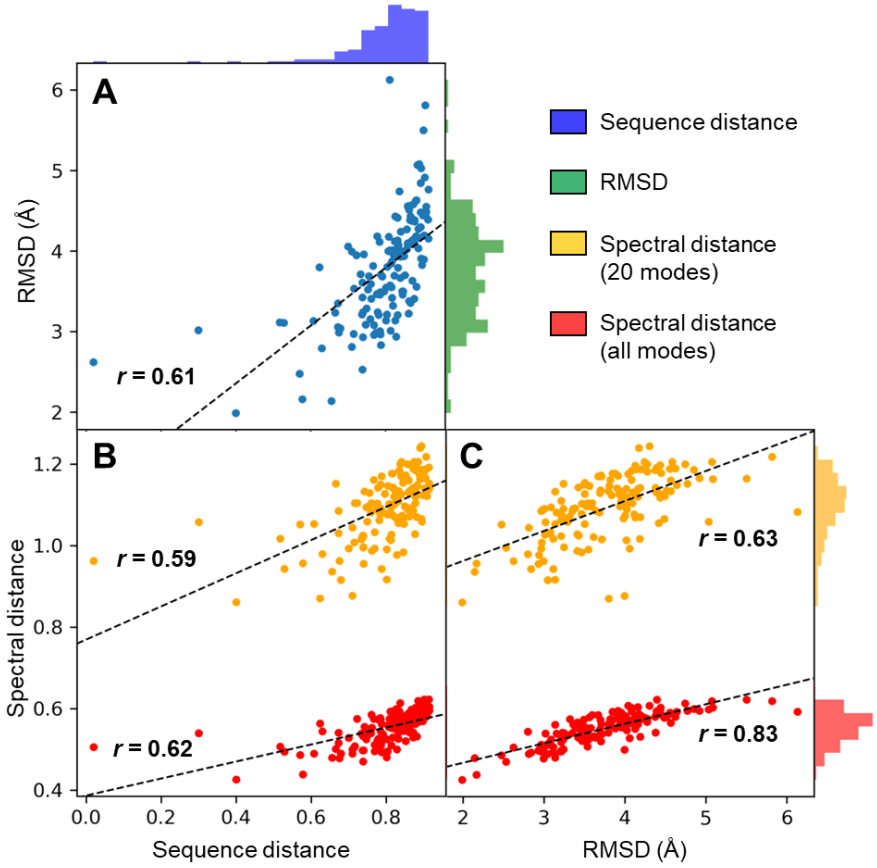

Figure 1.12 Correlations and distributions between sequence, structure, and dynamical (dis)similarities among members of 116 CATH superfamilies. (A) Correlation between average pairwise sequence distances and RMSDs among members of superfamilies (each *dot* represents a superfamily average). The histogram on the *top* and *right* shows the distribution of the average pairwise sequence distances and RMSDs of superfamilies, respectively. **(B)** Same result for the correlation between average pairwise spectral distances and sequence distances. **(C)** Same result for the correlations between average pairwise spectral distances and RMSDs. Histograms on the *right* show the distributions of the average pairwise spectral distances calculated using first 20 modes (*yellow*) or all modes (*red*).

1.2.2.4 Motions in the low-to-intermediate frequency regime differentiate the dynamics of family members

Figure 1.11A-C illustrates the spectral overlaps (*blue curves*) for the three example folds. In each case, the cumulative spectral overlap $\langle SO_k \rangle$ is plotted as a function of the total number of modes, *k*, included in the analysis, together with the corresponding variation among family members (*lighter blue band*). The curves reflect two counter-effects: First, there is a peak at the lowest-frequency end, consistent with the conservation of global modes. The overlap rapidly decreases with increasing *k*, due to the dissimilarity of the newly added modes. This differentiation between family members is consistent with the rapid drop in mode conservation shown in **Figure 1.11A-C** for LeuT, PBP-1, TIM barrel families as well as that for CATH superfamilies (**panel D**). Then, a new regime is observed, the low-to-intermediate frequency (LTIF) regime, which includes modes 20 to 60 approximately, where the spectral overlap is minimized. Finally, an opposite effect takes over, manifested by an increase in overlap. This arises from the increased coverage of the space of conformational changes (shown in the *orange curve*), consistent with the theoretical limit of $SO_k\ (A, B) \rightarrow 1$ as the complete space of motions is considered. The minimum in $\langle SO_k \rangle$ occurs for $k \leq 50$.

The LTIF regime where the cumulative spectral overlap is minimized emerges as a determinant of the specificity of family members. The percent contribution of the modes in this regime to the overall spectrum amounts to ~25% (see the increase in the cumulative weight of modes (*orange curves* in **Figure 1.11A-D**) in this interval), which means a substantial contribution to alter dynamics, while retaining the generic behavior.

Further calculations performed for CATH superfamilies (**Figure 1.11D**) corroborated the same trends. Supplementary Table S4 in (Zhang, et al., 2019) lists the spectral overlap calculated

for k = 3, 20 and all n - 1 modes for each superfamily, along with their standard deviations, and **Figure 1.13** displays their histogram. The spectral overlap achieved by global modes, $\langle SO_3 \rangle$, averaged over all superfamilies, is 0.55 ± 0.25, despite the low (< 0.10) cumulative weight of this small set of modes. The addition of modes in the LF regime lowers the cumulative overlap to 0.45 ± 0.15, even though a larger subspace of conformational changes is sampled, indicating the dissimilarities in conformational motions among members in this regime. A high overlap ($\langle SO_{all} \rangle$ = 0.84 ± 0.02) is recovered by the ensemble of all modes, which, by definition, forms a complete basis set that spans all possible conformational changes.

Overall, these data underscore the role of motions in the LTIF regime in differentiating family members within a given fold family, which will be further elaborated below.

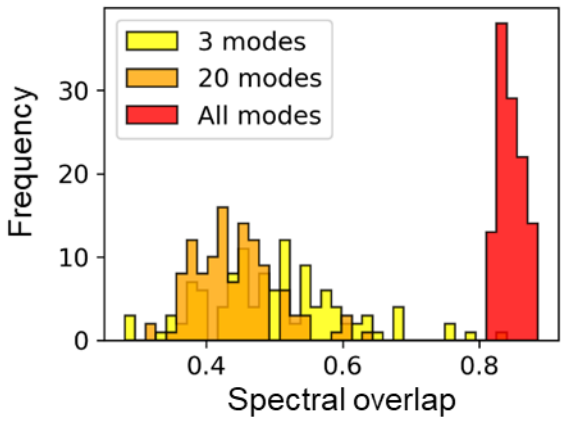

Figure 1.13 Distributions of average pairwise spectral overlaps between members of 116 CATH superfamily members. Results were evaluated based on global ($k \leq 3$), HF ($k \leq 20$), and all ($k \leq n$-1) modes.

1.2.2.5 Increased sequence heterogeneity among the members of a given fold family manifests itself by higher differentiation of dynamics

A previous study showed that sequentially conserved sites are also distinguished by their restricted fluctuations; or the mobility of residues, reflected by their root-mean-square fluctuations

(RMSFs) around their mean positions, increases with increasing Shannon entropy (H) at the corresponding sequence position (Liu and Bahar, 2012). That study established the correlation between sequence variation and conformational flexibility (RMSF). Here we investigated one further property, the change in flexibility, ΔRMSF, at a given position among family members, which is a metric of the extent of differentiation in the equilibrium dynamics between family members.

To this aim, we first evaluated the level of sequence heterogeneity within each family, using Shannon entropy as a metric. The resulting distribution among 13,648 residues belonging to 77 CATH families (after excluding the small folds with $n < 100$ residues) is shown by the histogram (*gray bars*) in **Figure 1.11E**. The histogram perfectly fits a lognormal distribution in support of the accurate sampling of sequence variabilities by the examined set. The changes in residue fluctuations, $\Delta\langle RMSF\rangle$ (where the triangular brackets indicate the averages over residues with sequence entropy in the bin corresponding to the *bar* underneath), exhibit a smooth increase with increasing sequence entropy (four *curves* in **Figure 1.11E**), confirming that sequentially diverse families exhibit higher differentiation in their dynamics.

The results are presented for different subsets of modes: global ($k \leq 3$), LF ($4 \leq k \leq 20$), LTIF ($21 \leq k \leq 60$), and HF ($k \geq 60$) regimes. The bar plot in the **Figure 1.11E** *inset* displays the $\Delta\langle RMSF\rangle$ averaged over all sequence entropies for the four respective groups. These results clearly show the dominant role of LTIF motions in imparting the member-specific differences in the fluctuation spectrum of individual family members, except for the high sequence entropy region. In this case, the differentiation of the modes shifts towards slower modes, as can be seen from the crossover between the LF and LTIF curves. The shift to LF modes reflects the earlier divergence of modes along the mode spectrum, in tandem with the higher divergence of sequence.

Overall, this analysis demonstrates that sequence divergence is accompanied by a divergence in structural dynamics, even though the fold is conserved.

A closer examination shows that $\Delta\langle RMSF\rangle$ contributed by the global modes is relatively flat with respect to sequence entropy ≤ 1.5. This insensitivity to sequence variations suggests that global motions are more conserved compared to sequence, presumably consistent with the slower divergence of structure, compared to sequence. **Figure 1.12C** further shows that diverging structures encode diverging dynamics despite the rather narrow root-mean-square deviation (RMSD) range. This dependency is stronger when all modes (*red dots*) are considered, as opposed to global modes (*orange dots*), confirming the increased differentiation of mode spectra with addition of higher modes. There is, however, some variation of spectral overlap with sequence identity (**Figure 1.12B**), again confirming that diverging sequences encode diverging dynamics as well.

1.2.2.6 Differentiation of protein families into specific subfamilies is accompanied by the evolution of LTIF motions

Consider a family composed of *m* subfamilies (or a superfamily of *m* families). For example, the currently considered TIM family contains 8 subfamilies (with at least 4 members). Subfamily classification is based on the specific functions of family members, e.g. in the case of TIM barrel, we have aldolases class 1 (ALD1), glycosidases (GLYC), phosphoenolpyruvate binding domains (PEPE), etc. Of interest is to assess to what extent subfamily members share similar modes among themselves, and to what extent they differ from other subfamily members. In other words, is the differentiation of fold families into specific subfamilies accompanied, if not driven, by a subset of modes that typifies the subfamily, and distinguishes it from all other subfamilies?

Note that subfamily members are not necessarily sequentially close or structurally close, but they belong to the same subfamily because of their shared biological (e.g. specific enzymatic) activities. In this respect, it is of interest to see if their common functions are supported by common mechanisms of action, or shared modes. Another way of asking the same question is which particular modes, or modes in which frequency regime, unify members within subfamilies, while ensuring maximal differentiation between subfamilies themselves. Toward this goal, we evaluated the spectral distances $\langle d_{ij} \rangle_{m_p,m_s}$ between subfamilies *p* and *s*, composed of m_p and m_s members respectively, based on the similarity of their modes $i \leq k \leq j$ (see Materials and Methods and Supplementary Information).

Figure 1.14 illustrates the respective results for TIM families. Results are presented for the global, LF, LTIF and HF regimes (respective **panels A-D**). The maps in each case are color-coded by the distances between the dynamics of the subfamilies listed along the two axes (see the color-code on the *right*). Note that the diagonal elements describe the level of conservation of dynamics within subfamilies (averaged over all combinations of family members); whereas off-diagonal terms represent the distances between pairs of subfamilies, with dark red entries indicating a strong divergence. We note that the LTIF modes are maximally distinctive across families, followed by LF modes, while the global modes and, interestingly, HF modes (>60) retain similarities. The strong discrimination provided by the LTIF regime between subfamilies - a feature apparent in the large-scale examination of CATH superfamilies, is now clearer with the subfamily-subfamily distance maps based on subfamily dynamics.

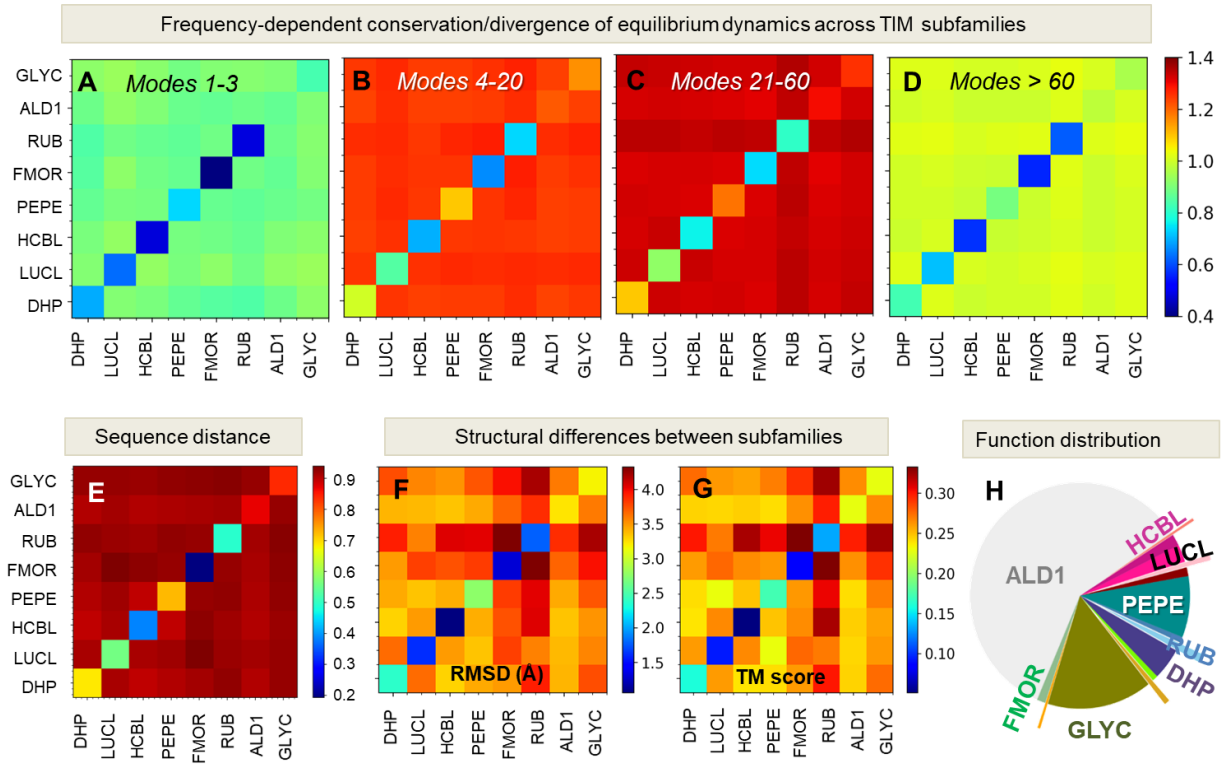

Figure 1.14 Low-to-intermediate frequency (LTIF) modes discriminate between subfamilies with different functions belonging to the TIM barrel fold family. (A-D) Subfamily-subfamily distance matrices based on structural dynamics, evaluated for eight TIM subfamilies. Subfamily acronyms are listed along the axes (see full names in **Figure 1.5**). Spectral distances $\langle \boldsymbol{d_{ij}} \rangle_{\boldsymbol{m_p},\boldsymbol{m_s}}$ averaged over all $\boldsymbol{m_p}$ and $\boldsymbol{m_s}$ members of respective subfamilies are shown by color-coded elements (*red*: long; *blue*: short; see the *bar* on the *right*). Results are displayed for four frequency regimes, global, LF, LTIF and HF, in the respective panels A-D, as indicated by the ranges $i \leq$ mode $\leq j$. Diagonal terms show the distances between within subfamilies based on the motions in the particular frequency window; and the off-diagonal terms show those across subfamilies. The LTIF regime (modes 21-60) provides the sharpest discrimination between subfamilies; whereas modes in both the global **(A)** and high-frequency **(D)** regimes are relatively conserved. For comparison, we present the sequence distances **(E)** and structural distances (**E** and **G**, using RMSD and TM-score as metrics) between subfamilies. Note that the subfamily-subfamily spectral distances in the LTIF regime (panel C) conform closely to their functional classification (panel E) defined by CATH, rather than their structural similarities (panels F-G), in strong support of the significance of LTIF motions in the evolution of function. This figure is adapted from (Zhang, et al., 2019).

Further comparison of the conservation/divergence of structural dynamics across subfamilies with their sequence and structure similarities (**panels E-G** in **Figure 1.14**) reveals that the correlations (or lack thereof) between the mode spectra of subfamilies in different regimes closely parallel sequence properties, rather than structural similarities/dissimilarities. The latter was assessed by two metrics, average RMSD between subfamilies and average TM-score (template modeling score)(Zhang and Skolnick, 2004), which yielded almost identical results. In other words, the division of families into subfamilies relates to the differentiation of their dynamics, more than the differentiation of their structure, in support of the direct relevance of motions/dynamics to subfamily function. The same analysis was repeated for PBP-1 superfamily and yielded similar results (data not show; see Supplementary Figure S4 in (Zhang, et al., 2019)). Overall, these results demonstrate that the specific mechanisms that distinguish subfamilies can be traced back to the intrinsic modes in the LTIF regime.

1.2.3 Discussion

Structural stability and related functions such as residue packing density are key constraints in sequence conservation and evolutionary change rate (Echave, et al., 2016). Yet, stability alone is not sufficient for functionality. Many proteins achieve their function by virtue of their conformational flexibility (Haliloglu and Bahar, 2015; Skjaerven, et al., 2011; Zheng, et al., 2009). While the conservation of sequence closely relates to structural stability and thermodynamics, the conservation of structure and its evolution might be closely determined by its adaptability to functional requirements. Pioneering studies that introduced the concept of evolution of structural dynamics and/or its relation to sequence evolution traditionally focused on experimental data, e.g. α-carbon fluctuations (B-factors) (Maguid, et al., 2008; Maguid, et al., 2006), the coupling between

sequence variability and structural dynamics (Liu and Bahar, 2012; Nevin Gerek, et al., 2013), or diversity of conformers resolved for well-studied proteins in the PDB (Juritz, et al., 2013). In this study, we applied our newly implemented interface, SignDy, to systematically analyze 15,636 proteins in 77 CATH superfamilies. Our analysis revealed features that could not be discerned if it were not for serial analysis of such large ensembles of CATH superfamilies. Decomposition of the mode spectrum into the contribution of different frequency windows permitted us to discern for the first time the differences in the conservation of modes in different frequency regimes, and elucidate the close relationship between the dissimilarities in the LTIF modes and the structural variations that distinguish subfamilies.

It is well known that sequence diverges much faster than structure. In other words, the sequence space is much larger than the structure/fold space. The mapping of various sequences into a small number of folds, or a relatively small set of fold superfamilies (e.g. ~100 examined here that cover ~ 20% PDB structures), does not, however, prevent proteins from achieving a broad diversity of functions. The latter is enabled by conformational dynamics, which is suggested by the present results that it supports the selection of folds in two ways: First, all family members share the fold-encoded global modes, or signature dynamics, that presumably underlie the versatility of the fold, e.g. the different members may exhibit different levels of inter-domain opening, or global twisting, but these are all slight rearrangements along the shared soft modes, which facilitate the adaptation to different substrates (Batista, et al., 2011; Batista, et al., 2010; Krieger, et al., 2015; Ponzoni, et al., 2018). Secondly, motions in the LTIF regime define the specificity of subfamilies. As a result, members of subfamilies are unified by their shared motions, or mechanisms of actions, in the LF regime, whereas they are differentiated from other subfamily members by virtue of the differences in their specific motions mainly in the LTIF regime.

Despite the wealth of data on well-studied proteins such as TIM barrel proteins, it is still not clear whether their shared fold originates from common ancestry, or results from convergent evolution. Protein folds are presumed to be more susceptible to evolutionary convergence than sequences, but sequence-profile based phylogenetic analysis can detect evolutionary relationships even among sequentially distant members of a given superfamily, in support of divergent evolution (Theobald and Wuttke, 2005). Other studies show that fitness constraints enforce evolutionary paths that preserve protein structure despite sequence divergence down to 30% sequence identity (Gilson, et al., 2017). Yet, the currently examined superfamilies contain members with much lower sequence identity, and other studies suggest that there is a limit to amino acid divergence while maintaining the contact topology/fold of the protein (Porto, et al., 2005).

While the current study cannot ascertain whether the shared structures are maintained during divergent evolution of sequences, or selected by convergent evolution, we clearly distinguish robust signature dynamics shared by family members, as well as LF and LTIF modes that characterize subfamilies. It remains to be established whether the prevalence of robust global motions, and accessibility to selected LTIF modes drive the selection of these folds. These challenges call for more research in comparing the differentiation of protein dynamics with sequence divergence, where SignDy can serve as an analytical framework for facilitating such investigations. For example, our recent examination of the signature dynamics of the lipoxygenase family of proteins and its differentiation among members helped detect the sites that enable its adaptation to specific substrate binding and allosteric activity (Mikulska-Ruminska, et al., 2019). We expect the SignDy interface to serve as a resource for efficiently analyzing family of proteins and designing allosteric modulators that can specifically target selected members of the family.

1.2.4 Methods

SignDy is an integrated pipeline for evaluating the signature dynamics of protein families based on ENMs. The pipeline comprises seven major steps (**Figure 1.15**):

1. **Selection of protein family.** The input to SignDy can be fed or generated in three ways: (i) entering a Pfam (El-Gebali, et al., 2019) or CATH (Dawson, et al., 2017; Knudsen and Wiuf, 2010) ID; (ii) providing a query PDB (Berman, et al., 2000) or UniProt (UniProt, 2019) ID, or sequence whose structural homologs are retrieved using from the Dali server (Holm and Laakso, 2016; Holm and Rosenstrom, 2010); or (iii) submitting a manually prepared list of homologous proteins.
2. **Selection of a representative set of homologs and the reference protein.** Overrepresented sequences and structures as well as highly dissimilar ones are filtered out based on default or user-selected thresholds for sequence or structural similarities. This leads to a family of *m* members, one of which is the reference structure, *R*, based on prior knowledge or pre-defined criteria.
3. **Structural alignment and definition of core atoms/residues.** Several alignment tools can be used to identify and align corresponding residues forming shared structural elements (core structure) with *n* core residues as described below.
4. **Evaluation of mode spectra**, using the GNM or ANM. We use a system-environment framework where the core is treated as the "system" and other residues as the "environment" (Hinsen, et al., 2000; Zheng and Brooks, 2005). To identify shared motions among family members, we examine the mode-mode correlations between *R* and each of the other *m* - 1 members. The resulting equivalent modes for each member are reordered to match the mode order of *R*.

5. **Analysis of signature dynamics**. The spatial mobility of core residues driven by global modes averaged over all members, and its variation across family members define the "signature dynamics" of the family. Other generic properties include the averages and deviations of cross-correlations between residue motions across family members. Departures from signature dynamics or cross-correlations highlight the specific features of individual members.
6. **Assessment of mode conservation and spectral overlap between family members**. The level of conservation of mode *k* among family members is measured by the correlation cosine computed for the k^{th} equivalent GNM mode of pairs of members, averaged over all $m(m-1)/2$ pairs (*green curves* in **Figure 1.11**). The overall extent of similarity between the mode spectra of members *A* and *B*, based on the *k* softest modes, is given by the spectral overlap (see Section 1.2.4.5).
7. **Classification of family members based on their dynamics**. A dynamics-based dendrogram for each family can be constructed using the spectral distance between pairs of family members as a metric (see Section 1.2.4.6).

The above procedures were implemented in *ProDy* (Bakan, et al., 2014; Bakan, et al., 2011). More details are presented in the following subsections.

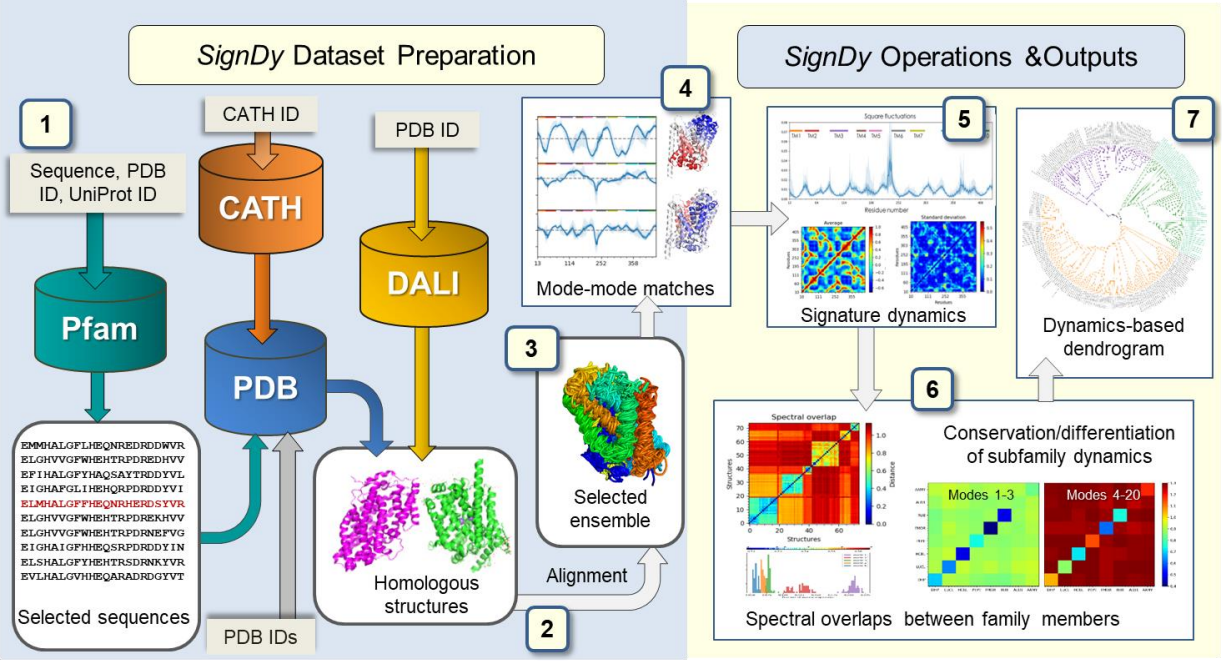

Figure 1.15 SignDy workflow. The workflow is separated into two main parts: dataset preparation (*left*; steps 1-3) and SignDy operations and outputs (*right*; steps 4-7), described in the text above. *Cylinders* and *light grey rectangular boxes* represent databases and corresponding query inputs, respectively. This figure is adapted from (Zhang, et al., 2019).

1.2.4.1 Dataset of CATH Superfamilies

We considered the 175 most populated superfamilies in the CATH database (Dawson, et al., 2017; Knudsen and Wiuf, 2010), and selected 116 comprised of a total of 26,899 proteins after eliminating the close structural homologs (RMSD < 1 Å) using single-linkage clustering, as well as outliers (RMSD ≥ 10 Å with respect to the reference), and superfamilies with less than 50 representative members. We superposed all the members to their reference in each superfamily using the CE algorithm (Shindyalov and Bourne, 1998). Supplementary Table S4 in (Zhang, et al., 2019) lists all the superfamilies and their properties.

1.2.4.2 Datasets for LeuT, PBP and TIM barrel fold families

LeuT. The structural ensemble is prepared as described in Section 1.1.4.1.

PBP-1. Individual PBP-1 domain structures were selected and aligned with Dali server using a structure of the N-terminal domain (NTD) of AMPA-type ionotropic glutamate receptor (iGluR) paralogue GluA2 (chain A from PDB ID: 3H5V) (Jin, et al., 2009) as the query. This search yielded a set of 2,291 chains from 977 structures including isolated domains and whole receptors. We filtered out those results with Dali Z-score below 10 and less than 50% coverage, resulting in 971 chains from 451 structures. We further refined the set by applying an RMSD filter that removed redundant structures within 1.0 Å RMSD from others in the ensemble, as well as the outliers (≥10.0 Å RMSD from all others). This led to an ensemble with 379 members (data not shown; see Supplementary Table S2 in (Zhang, et al., 2019)), including iGluR NTDs, class C G-protein coupled receptor (GPCR) and natriuretic peptide receptor ligand-binding domains, and bacterial periplasmic binding proteins (PBPs) and transcription regulators (TRs). **Figure 1.5** displays the histograms of pairwise sequence identities, structural RMSDs and biological function for PBP family members.

TIM barrels. TIM barrel structures were selected and aligned by Dali using the triose phosphate isomerase (TIM) structure with PDB ID 8TIM (chain B) (Banner, et al., 1975) as the query. The search yielded a total of 1,070 structures. Among them, 455 were filtered out by requiring the following criteria to be satisfied: RMSD > 1 Å with respect to the query structure; Dali Z score > 10; and coverage > 70%. Among the remaining 615 structures, 14 could not be aligned using the mapping information from Dali, which led to 601 structures. As an additional filter, we excluded members from all pairs outside the range 1 < RMSD < 10 Å. This led to an ensemble of 290 conformations and the columns in the multiple sequence alignment (MSA) were

trimmed to ensure column occupancies of 0.7 or higher, resulting in 180 columns corresponding to core residues (data not shown; see Supplementary Table S3 in (Zhang, et al., 2019)).

1.2.4.3 Calculation of mode spectra and sorting of modes

GNM or ANM analyses are performed for each member, using its complete structure composed of the core (shared by all members) and other residues (specific to members) using the system-environment framework (Hinsen, et al., 2000; Zheng and Brooks, 2005). The effect of the environment on the core dynamics is modelled therein by adopting a modified Kirchhoff (GNM) or Hessian (ANM) matrix for the core. In this way, we identify the mode spectrum for each family member. The high-throughput examination of protein family dynamics is possible because of the efficiency of ENMs.

Because of the structural variations, the order (or relative frequencies) of the modes may vary among family members. Pairwise comparisons of the mode spectra of family members necessitate the identification of the equivalent modes. We accomplish *mode-mode matching* as a linear assignment problem (Kuhn, 1955; Kuhn, 1956). Accordingly, we first calculate the correlation cosine, $\rho_{ij}(A,B) = \boldsymbol{v}_i^{(A)} \cdot \boldsymbol{v}_j^{(B)}$, between each pair of modes i and j belonging to proteins A and B, then evaluate the cost of matching them as $1 - \rho_{ij}(A,B)$, and finally select the set of pairs that minimizes the total cost.

1.2.4.4 Evaluation of signature dynamics

The signature dynamics is defined by global dynamics shared by family members. It refers to any dynamical property of a protein structure that can be derived from the ENM which is now evaluated for every conformer in the structural ensemble and aligned altogether. Such dynamical

properties mainly involve three: 1) modes of motions; 2) the MSFs of residues driven by a selected subset of global modes or all modes; 3) the cross-correlations between residue fluctuations. The averages of these properties over all members, respectively denoted by $\langle \boldsymbol{v}_k \rangle$, ⟨MSF⟩ and $\langle \mathbf{C} \rangle$, describe the *generic* behavior of the family. Note that each mode describes a fully symmetric fluctuation; so $\boldsymbol{v}_k$ and $-\boldsymbol{v}_k$ represent the same eigenvector for mode *k*; and therefore eigenvectors are assigned the same sign as their counterparts in the reference structure before evaluating the averages $\langle \boldsymbol{v}_k \rangle$ over family members.

The departures of the individual members from the generic behavior are given by the standard deviations $\Delta \boldsymbol{v}_k$, ΔMSF and $\Delta \mathbf{C}$, displayed by, for example in **Figure 1.6**, a band around the mean values ($\Delta \boldsymbol{v}_k$ and ΔMSF) or by an additional $n \times n$ map ($\Delta \mathbf{C}_{ij}$).

1.2.4.5 Spectral overlap and mode-mode overlap

We use the spectral overlap (termed as covariance overlap in (Hess, 2002)) as a measure of the degree of similarity between the global mode spectra of structures *A* and *B*. The spectral overlap provides a robust and easy-to-compute metric, as a function of the entire set of eigenvalues and eigenvectors, to measure the overlap of the subspaces spanned by two mode spectra or subsets of modes. The spectral overlap based on *k* LF modes (i.e. mode index in the range [1, *k*]) predicted by the ENM is defined as

$$SO_k(A,B) = 1 - \left[\frac{\sum_{i=1}^{k} (\sigma_i^{(A)} + \sigma_i^{(B)}) - 2\sum_{i=1}^{k}\sum_{j=1}^{k} (\sigma_i^{(A)} \sigma_j^{(B)})^{\frac{1}{2}} (\boldsymbol{v}_i^{(A)} \boldsymbol{v}_j^{(B)})^2}{\sum_{i=1}^{k} (\sigma_i^{(A)} + \sigma_i^{(B)})} \right]^{\frac{1}{2}}, \quad (1.8)$$

where $\sigma_i^{(A)}$ designates the i^{th} eigenvalue of the covariance matrix for protein *A*, and $\boldsymbol{v}_i^{(A)}$ is the corresponding eigenvector. Note that $\lambda_i^{(A)} = 1/\sigma_i^{(A)}$, where $\lambda_i^{(A)}$ is the i^{th} eigenvalue of the connectivity matrix (Hessian for the ANM; Kirchhoff for the GNM) for same protein *A*.

$SO_k(A,B)$ varies in the range [0, 1]. The upper limit can be only reached by two entirely overlapping mode spectra. For each superfamily and a given *k*, the spectral overlap $\langle SO_k \rangle$ is averaged over all $\frac{m(m-1)}{2}$ pairs of *A* and *B*.

A detailed analysis of the extent of differentiation between the individual modes of family members is performed by evaluating the *correlation cosines*, also called *mode-mode overlaps*, averaged over all $\frac{m(m-1)}{2}$ pairs

$$\langle cc_k \rangle = \frac{2}{m\,(m-1)} \sum_A \sum_{B \neq A} \left| \boldsymbol{v}_k^{(A)}\, \boldsymbol{v}_k^{(B)} \right|. \tag{1.9}$$

The absolute value of the correlation cosine is used because the direction of the eigenvectors is immaterial. Note that the mode number *k* refers to the rank-ordered index determined after identifying the optimal matches between the mode spectra of family members, as described in the Section 1.2.4.3.

1.2.4.6 Spectral distance and construction of dynamics-based dendrograms

The spectral distance, $D_k(A,B)$, between the first *k* global modes of *A* and *B* is defined by the arc cosine $D_k(A,B) = \cos^{-1}(SO_k(A,B))$ and that among all members of a given family is evaluated as

$$\langle D_k \rangle = \frac{2}{m\,(m-1)} \sum_A \sum_{B \neq A} \cos^{-1}[SO_k(A,B)]\,. \tag{1.10}$$

The $m \times m$ distance matrix $D_k(A,B)$ with $k = 3$ is used as metric for classifying family members based on their global mode spectra. The dendrograms (**Figure 1.9**) are constructed using the neighbor joining (NJ) (Saitou and Nei, 1987) or Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sokal, 1958) method. Similar trees based on sequence and structure dissimilarities allow for comparing the differentiations of sequence, structure and dynamics among

the members of the family. Here we adopted the RMSDs after structural alignment as structure distance, and the Hamming distance $D_H(A, B)$ (normalized by the number of columns in the MSA) as sequence distance between members *A* and *B*.

1.2.5 Acknowledgment

The present work was published (Zhang, et al., 2019) with me being one of the first coauthors. Drs. Hongchun Li and James Krieger (the other two first co-authors) made equally significant contributions. H.L. implemented the interface to Dali server and CATH database in ProDy and performed the analysis on TIM barrels. J.K. implemented part of the SignDy functions and performed the analysis on PBP domains. I implemented the SignDy interface (as a module of ProDy) and performed the analysis on LeuT folds. H.L, J.K. and I performed the analysis on CATH superfamilies. Dr. Ivet Bahar conceived the research plan and supervised the project.

2.0 Chromatin Dynamics Analyzed by the Gaussian Network Model

The spatial organization of the eukaryotic genome has emerged as a topic of broad interest over the past decade, as a consequence of its inherently complex structure emerging from recent studies, and its important role in regulating transcriptional activity (Bickmore and van Steensel, 2013; Cavalli and Misteli, 2013; Fraser, et al., 2015; Hou, et al., 2012). However, due to the size and complexity of the genome structure (or ensemble of conformers sampled under physiological conditions), it is impractical to use traditional structure determination techniques such as crystallography or nuclear magnetic resonance (NMR). As alternatives, the genome structure is usually studied by examining the 3D contacts of genomic loci via experiments such as Fluorescent in situ hybridization (FISH) or Chromosome Conformation Capture (3C). Improved by the next generation sequencing (NGS) advances, the Hi-C technology scaled up the 3C technique to all-to-all and genome-wide detection of chromatin contacts, thus providing a more complete view of the entire genome for human and other species (Dixon, et al., 2012; Lieberman-Aiden, et al., 2009; Rao, et al., 2014; Stevens, et al., 2017). These studies revealed that the spatial organization of the genome is largely compartmental and hierarchical (**Figure 2.1**). In parallel, many computational methods have been developed and contributed to these and other characterizations of chromosomal architecture (Forcato, et al., 2017; Lajoie, et al., 2015; Oluwadare, et al., 2019). In particular, it was identified that chromosomal spatial regions are hierarchically categorized as nested structures of varied sizes, such as A/B compartments and topologically associating domains (TADs). However, the scale, complexity, and noise inherent in the available data still make it challenging to determine the exact spatial relationships and underlying chromatin architecture, let alone

structure-based dynamics. Therefore, the exact nature of chromatin spatial organization and its influence on gene expression and regulation remain unclear.

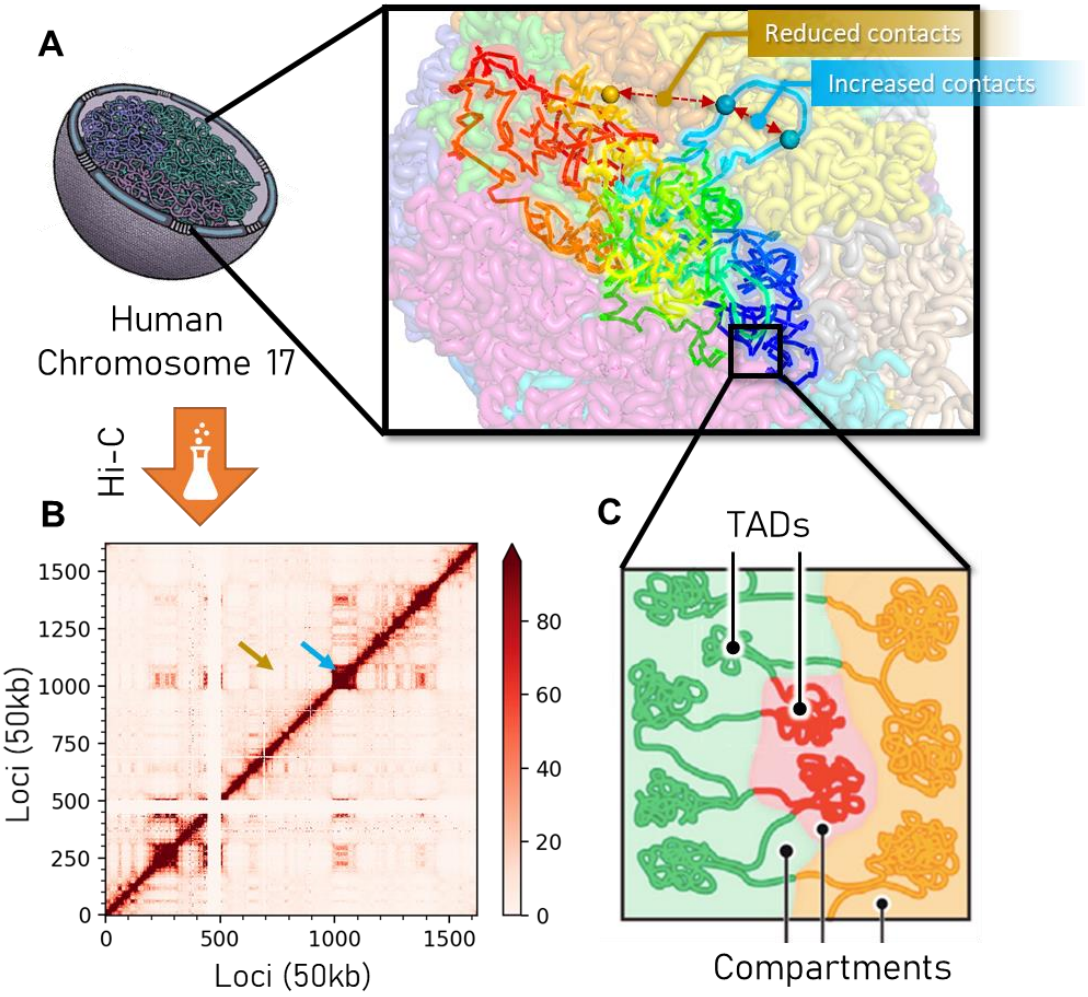

Figure 2.1 Spatial organization of mammalian genome (schematic). (A) Cell nucleus containing chromosomes in the form of chromatin fibers (*left*) and computationally resolved 3D model for manmalian chromosomes (Stevens, et al., 2017) (*right*). Three loci are highlighted and represented by *nodes*, and the spatial distances between them are indicated by *red dashed double-headed arrows*. **(B)** Chromatin contact map determined by Hi-C experiment. The spatial distance between the two *cyan* loci in panel A is smaller than that between the *orange* and the *cyan* ones, so the former is subject to a higher frequency (*cyan arrow*) of contacts than the latter (*yellow arrow*). **(C)** Two main levels of the chromosomal domains (Rao, et al., 2014): compartments (~5Mb) and topologically-associating domains (TADs, ~1Mb) (Rowley and Corces, 2018), manifested in the "blocks" along the diagonal of the Hi-C contact map (panel B).

In this chapter, we show that the GNM adapted to modeling chromatin inter-loci contact topology using Hi-C data provides a mathematically well-founded unified framework for modeling chromatin dynamics, assessing the structural basis of genome-wide observations, and identifying hierarchical chromosomal domains.

2.1 GNM Evaluation of Chromosomal Dynamics Explains Genome-Wide Accessibility and Long-Range Couplings

2.1.1 Introduction

The Gaussian Network Model (GNM) is a highly robust and widely tested framework developed for modeling the intrinsic dynamics of biomolecular structures (Bahar, et al., 1998; Bahar, et al., 1997; Bahar, et al., 2010). We adapted the GNM to the topology-based modeling of chromosomal dynamics. Chromosomal dynamics refers to the coupled spatial movements of loci under equilibrium conditions, as uniquely defined by the topology of an elastic network representative of the chromosome architecture. The only input GNM requires is a map of 3D contacts between structural elements that define the nodes of the network. Here, this information is provided by Hi-C data, which gives contact frequencies between genomic loci (network nodes). The Hi-C matrix is used for constructing the so-called Kirchhoff (or Laplacian) matrix which uniquely defines the equilibrium dynamics of the network, including the mean-square fluctuations of the nodes (genomic loci) as well as the spatial cross-correlations between pairs of nodes.

The chromatin structure is often described in terms of TADs, whose identification involves searching for sequentially contiguous groups of highly interconnected loci along the diagonal of

the Hi-C matrix of intra- chromosomal contacts. Spatial couplings between sequentially distant genomic regions, on the other hand, represent a new dimension to search and identify meaningful structural components. The identification of such long-range couplings is a challenging problem. Several methods have sought to identify long-range interactions from 3C-based data (Jin, et al., 2013; Rao, et al., 2014; Sanyal, et al., 2012; Xu, et al., 2016), but the scale of these interactions is still small compared to that of the full chromosome. Most methods detect interactions within 1–2 megabases (Mb), or up to 10 Mb (Ay, et al., 2014; Forcato, et al., 2017), so extending the span of predicted long-range couplings to the order of tens of millions of base pairs may yield new insights into regulatory actions. Such long-range correlations may originate from physical proximity in space, or other indirect effects similar to those in allosteric structures. Assessment of such long-range correlations is important for gaining a better understanding of the physical basis of gene expression and regulation.

We show and verify upon comparison with an array of experimental data and genome-wide statistical analyses that the GNM provides a robust description of accessibility to the nuclear environment as well as co-expression patterns between gene-loci pairs separated by tens of megabases. The analysis may serve as a framework for drawing inferences from Hi-C or other advanced genome-wide studies toward establishing the structural and dynamic bases of regulation.

2.1.2 Results

We evaluated the mobility profiles (MSFs) and covariance maps for the coupled movements of gene loci for GM12878 cells from a human lymphoblastoid cell line with relatively normal karyotype. We illustrate the results for chromosome 17 based on Hi-C data at 5 kb resolution (see Section 2.1.4 and **Figure 2.8**), based on Hi-C data at 5 kb resolution for GM12878

cells from a human lympho-blastoid cell line with relatively normal karyotype. We compared our predictions with the experimental measures of chromatin accessibility and interactions, namely DNase-seq (Tsompana and Buck, 2014), ATAC-seq (Buenrostro, et al., 2013), and Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) (Heidari, et al., 2014). We also examined co-expression enrichment for GNM-predicted cross-correlated distal domains (CCDDs), using RNA-seq expression data.

2.1.2.1 Correlation between chromatin mobility and experimental measures of accessibility

Figure 2.2 illustrates the MSFs obtained with the GNM (*blue curves*) for the loci on three chromosomes (1, 15 and 17, in respective **panels A**, **B** and **C**). Results for all other chromosomes are presented in Supplementary Figure S1 in (Sauerwald, et al., 2017).

GNM application to H/D exchange data (Bahar, et al., 1998) has shown that the MSFs of network nodes can be directly related to the accessibility of the corresponding sites: exposed sites enjoy higher mobility, and those buried have suppressed mobilities. The entropic cost of exposure to the environment for a given site can be shown to be inversely proportional to its MSFs based on simple thermodynamic arguments applied to macromolecules subject to Gaussian fluctuations (such as those represented by the GNM) (Bahar, et al., 1998). We examined whether GNM-predicted mobility profiles were also consistent with data from chromatin accessibility experiments. We compared our predictions with two measures of chromatin accessibility, DNase-seq (Tsompana and Buck, 2014) and ATAC-seq (Buenrostro, et al., 2013), shown respectively by the *yellow* and *red curves* in **Figure 2.2A-C**.

Figure 2.2 shows that the GNM-predicted MSFs of chromosomal loci are in good agreement with the accessibility of loci as measured by DNase-seq. The corresponding Spearman correlations for the three chromosomes illustrated in **panels A-C** vary in the range 0.78-0.85 (see

inset), and the computations for all 23 chromosomes (**panel D**, *yellow bars*) yield a Spearman correlation of 0.800 ± 0.044. The Spearman correlation between GNM MSFs and ATAC-seq data is somewhat lower: 0.552 ± 0.112. Interestingly, the Spearman correlation between the two sets of experimental data was 0.741 ± 0.089, suggesting that the accuracy of computational predictions is comparable to that of experiments, and that the DNase-seq provides data more consistent with computational predictions. ATAC-seq maps not only the open chromatin, but also transcription factors and nucleosome occupancy (Buenrostro, et al., 2013), which may help explain the observed difference.

We performed the same analysis on the available data for a different cell type, IMR90, and found an even better agreement with experiments (data not shown; see Supplementary Figure S2 in (Sauerwald, et al., 2017)). The Spearman correlation between the computed MSFs and experimental ATAC-seq data averaged over all chromosomes was 0.63 ± 0.08 IMR90 cells, and that between MSFs and DNase-seq data was 0.82 ± 0.03. Consistently, the two sets of experiments also exhibit a higher correlation (0.81 ± 0.06) in this case.

Overall, this analysis shows that GNM representation of chromosomal architecture using Hi-C data provides quantitative description of the spatial mobility of individual nodes (gene loci) in good agreement with their accessibility detected in DNase-seq and ATAC-seq experiments, the agreement with DNase-seq data being particularly strong. The same behavior is reproduced for all chromosomes in two different cell types.

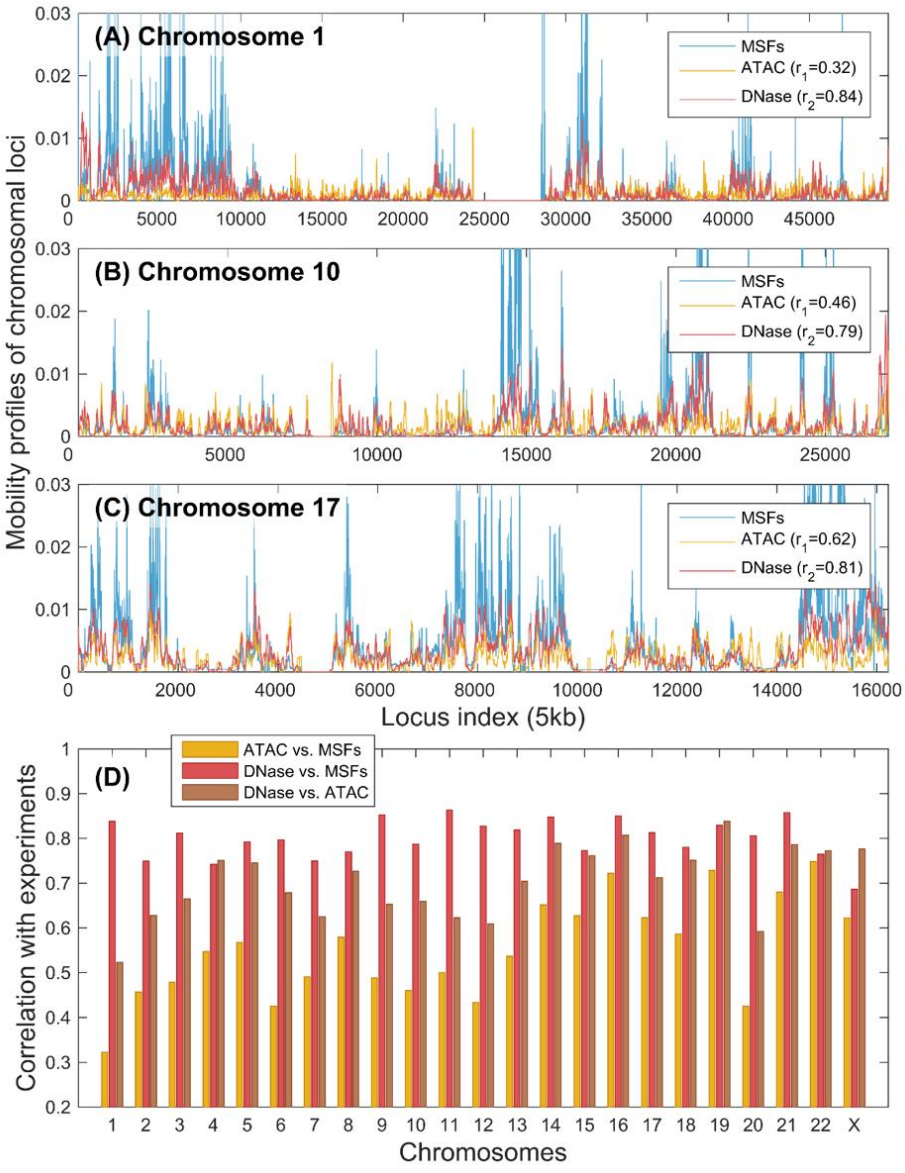

Figure 2.2 GNM-predicted mobilities of chromosomal loci in GM12878 show good agreement with data from chromatin accessibility experiments. (A-C) Mobility profiles (MSFs of loci, *blue*) obtained from GNM analysis of the equilibrium dynamics of chromosomes 1, 17, and X, respectively, are compared to the DNA accessibilities probed by ATAC- (*yellow*) and DNA-seq (*red*) experiments. GNM results are based on 500 slowest modes. $\boldsymbol{r_1}$ is the Spearman correlations between GNM predictions and DNase-seq experiments; and $\boldsymbol{r_2}$ is that between GNM and ATAC-seq. **(D)** Spearman correlations between theory and experiments for all chromosomes. See (Sauerwald, et al., 2017) for more details.

2.1.2.2 Robustness of GNM results

These results for two different types of cell lines show that the mobility profiles predicted by the GNM for the 23 chromosomes accurately capture the accessibility of gene loci. The agreement with experimental data lends support to the applicability and utility of the GNM for making predictions on chromatin dynamics. The current results were obtained by using subsets of 500 GNM modes for each chromosome, which essentially yield the same profiles and the same level of agreement with experiments as those obtained with all modes (**Figure 2.3**). The use of a subset of modes at the LF end of the spectrum improves the efficiency of computations, without compromising the accuracy of the results.

Computations repeated for different levels of resolution (from 5 to 50 kilobases (kb) per bin) also showed that the results are insensitive to the level of coarse-graining (**Figure 2.4**), which further supports the robustness of GNM results. We note that all results are obtained by adopting the vanilla coverage (VC) normalization for Hi-C data. Computations repeated with two alternative normalization schema, square-root VC (Rao, et al., 2014) and Knight-Ruiz (Knight and Ruiz, 2013) normalization, showed a significant decrease in the level of agreement with experimental data regardless of the number of modes included in the GNM computations and the underperformance of these schema became particularly pronounced in the case of high resolution data, in support of the VC normalization adopted here (**Figure 2.7**; see Section 2.1.4.2 for more details).

This analysis demonstrates that the GNM results are robust to model parameters such as the number of selected modes. A small subset of global modes (e.g. a few hundreds) provide an adequate description of the gene loci fluctuation profile.

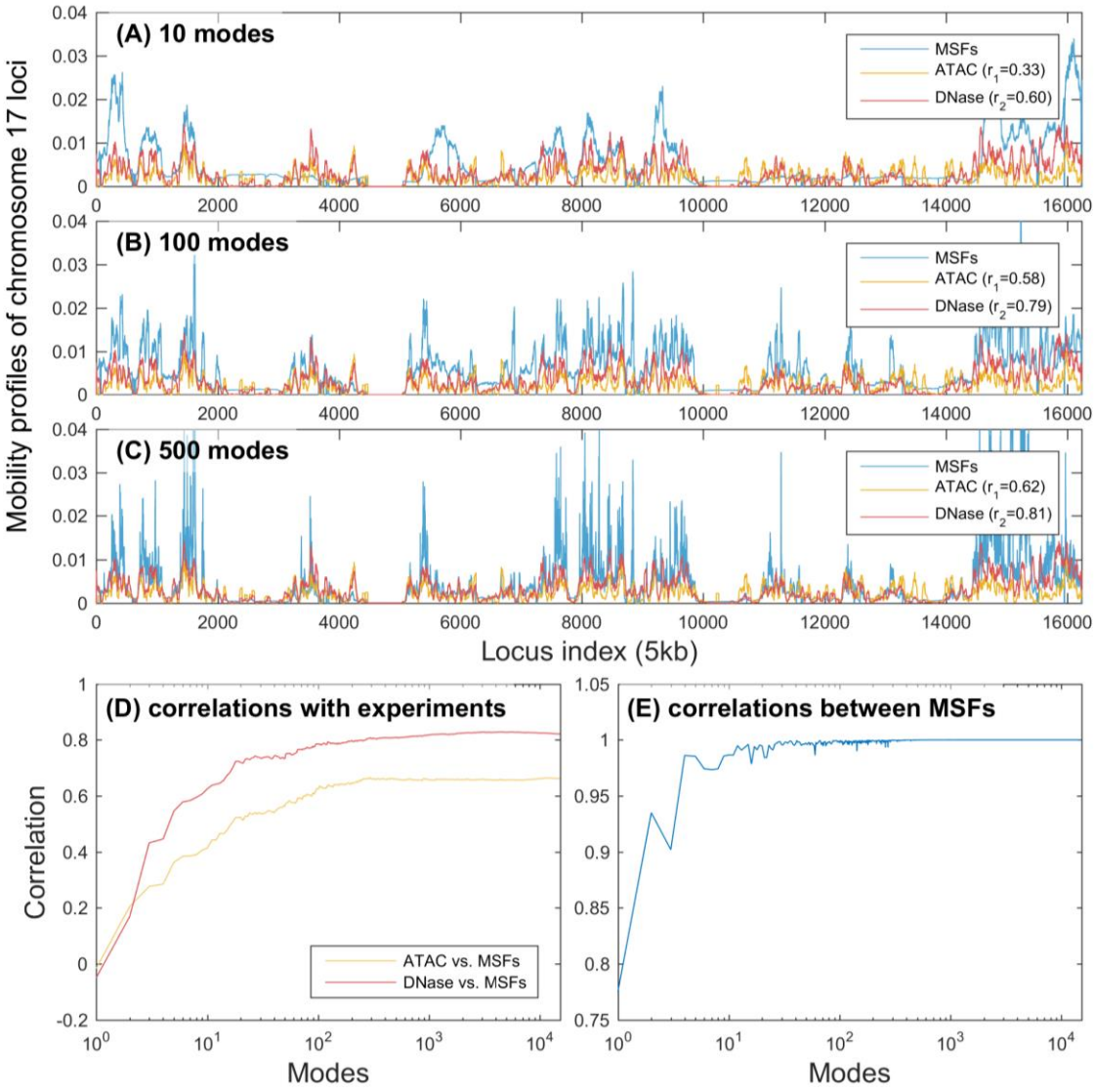

Figure 2.3 Mobility profiles computed using different subsets of GNM modes show the robust convergence of results with a small subset of modes. (A-C) Comparisons between experimental data and MSFs obtained using 10, 100, and 500 GNM modes. **(D)** Spearman correlations between experimental accessibility and computationally predicted mobility profiles obtained with different numbers of modes. **(E)** Spearman correlations between MSFs computed from slowest *k* modes and *k*+1 modes. *Abscissa* are in logarithmic scale in **panels D** and **E**. The correlation levels off at around a few hundreds of modes, showing that the addition of higher modes does not practically change the predicted MSFs, and a small subset of ~500 modes can be efficiently used for evaluating the MSFs (adapted from (Sauerwald, et al., 2017)).

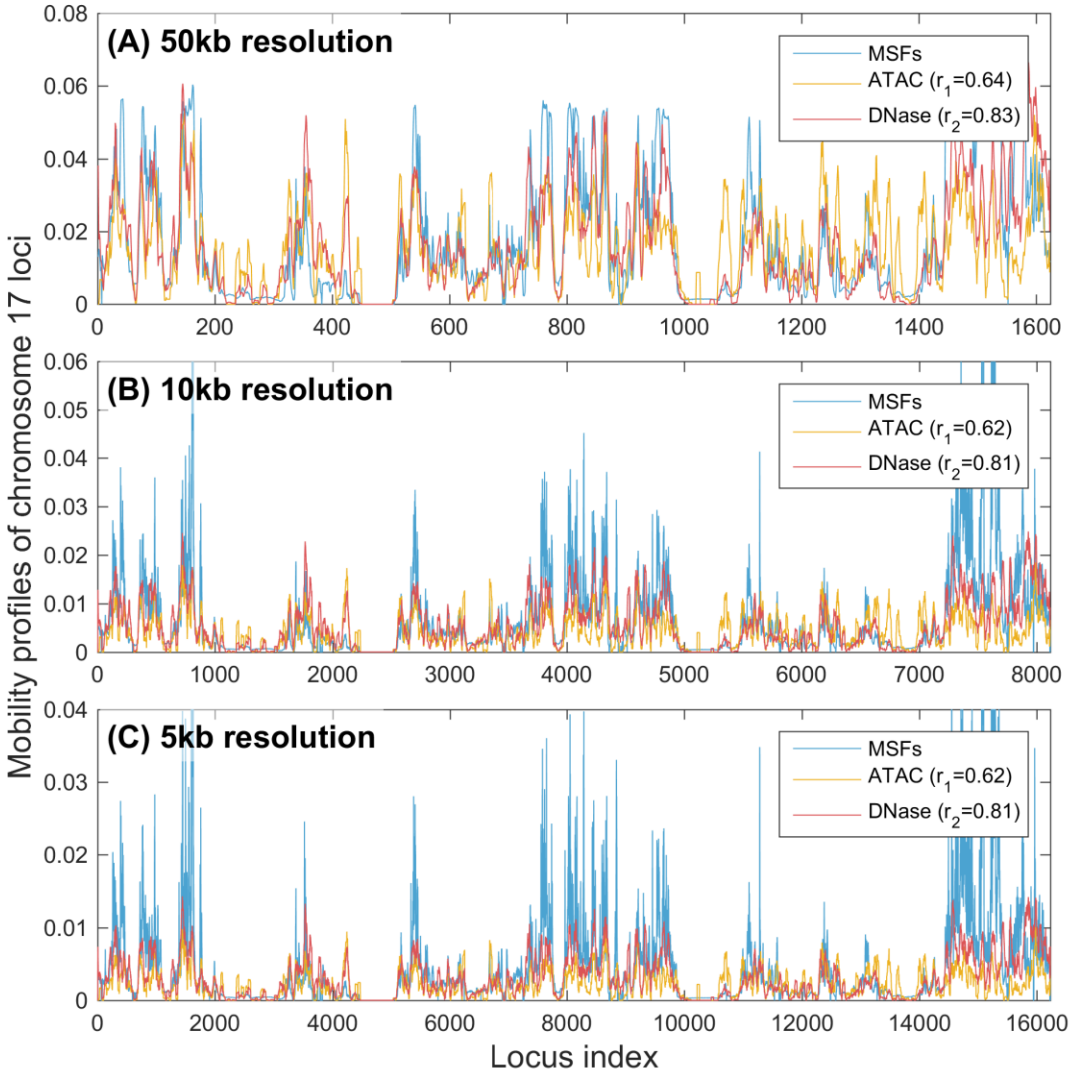

Figure 2.4 Mobility profile of GM12878 chromosome 17 predicted by the GNM based on Hi-C maps at different resolutions. The three panels display the correlations between chromatin accessibility data (ATAC and DNase-seq) and GNM-predicted fluctuation profiles based on the Hi-C contact map for chromosome 17 at **(A)** 50kb, **(B)** 10kb, and **(C)** 5kb resolution. GNM results are computed using 500 lowest-frequency modes. The level of agreement between computational predictions and experimental observations is insensitive to the resolution of experimental data. This figure is adapted from (Sauerwald, et al., 2017).

2.1.2.3 Loci pairs separated by similar genomic distances exhibit differential levels of dynamic coupling, consistent with ChIA-PET data

As presented in the previous Chapter, the GNM is a powerful tool for evaluating the cross-correlations between the movements of the network nodes. Here we examined what type of cross-correlations between the spatial displacements of gene loci would be predicted by the GNM. **Figure 2.5** displays the covariance map generated for the coupled movements of the loci on the chromosome 17 of GM12878 cells. **Panel A** displays the cross-correlations (see equation 2.5) between all loci-pairs as a heat map. Diagonal elements are the MSFs (presented in **Figure 2.2C**). The blocks along the diagonal (outlined by *dashed yellow boxes*) indicate loci of different sizes that form strongly coupled clusters. The *red dashed boxes* indicate the pairs of regions exhibiting weak correlations despite genomic distances of several megabases. The *curve* along the *upper abscissa* in **Figure 2.5A** shows the average cross-correlation of each locus with respect to all others; the peaks indicate the regions tightly coupled to all others, probably occupying central positions in the 3D architecture. Results for other chromosomes can be found in Supplementary Figure S12 in (Sauerwald, et al., 2017). The covariance map is highly robust and insensitive to the resolution of the Hi-C data. The results in **Figure 2.5A** were obtained using all 15,218 nonzero modes corresponding to 5 kb resolution representation of the chromosome 17. Calculations repeated with lower resolution data (50 kb) and fewer modes (500 modes) yielded covariance maps that maintained the same features (data not shown; see Supplementary Figure S13 in (Sauerwald, et al., 2017)).

Owing to their genomic sequence proximity, the entries near the main diagonal of the covariance map tend to show relatively high covariance values (colored *yellow-to-brown*; **Figure 2.5A**). Note that even the close vicinity of the diagonals (e.g. loci intervals of ≥200) represents (at

5 kb resolution) genomic loci separated by >1 Mb. The covariance map clearly shows that there are strong couplings between loci separated by a few megabases. We show an example of such regions in **Figure 2.5B**. While the loci pairs located in the *dark red band* along the diagonal appear all to exhibit strong couplings, a closer examination reveals differential levels of cross-correlations that are in good agreement with the data from Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) experiments (Heidari, et al., 2014). The "long-range" interactions identified by ChIA-PET are indicated in **panel B** by *red dots* (close to the diagonal). These are interacting loci separated by several hundreds of kb. We selected background pairs separated by the same 1D distance, on both sides of the ChIA-PET pair, and compared the cross-correlations predicted for the two sets along each chromosome (**Figure 2.5C**). The background pairs (*blue bars*) show weaker GNM cross-correlations compared to the ChIA-PET pairs (*red bars*) although they are separated by the same genomic distance along the chromosome.

Similar statistical analysis repeated for all 23 chromosomes showed that the cross-correlations between pairs of loci identified by ChIA-PET experiments were greater than those of background pairs separated by the same genomic distance on every chromosome, with all p-values being less than 10^{-19} (two-sided t-test).

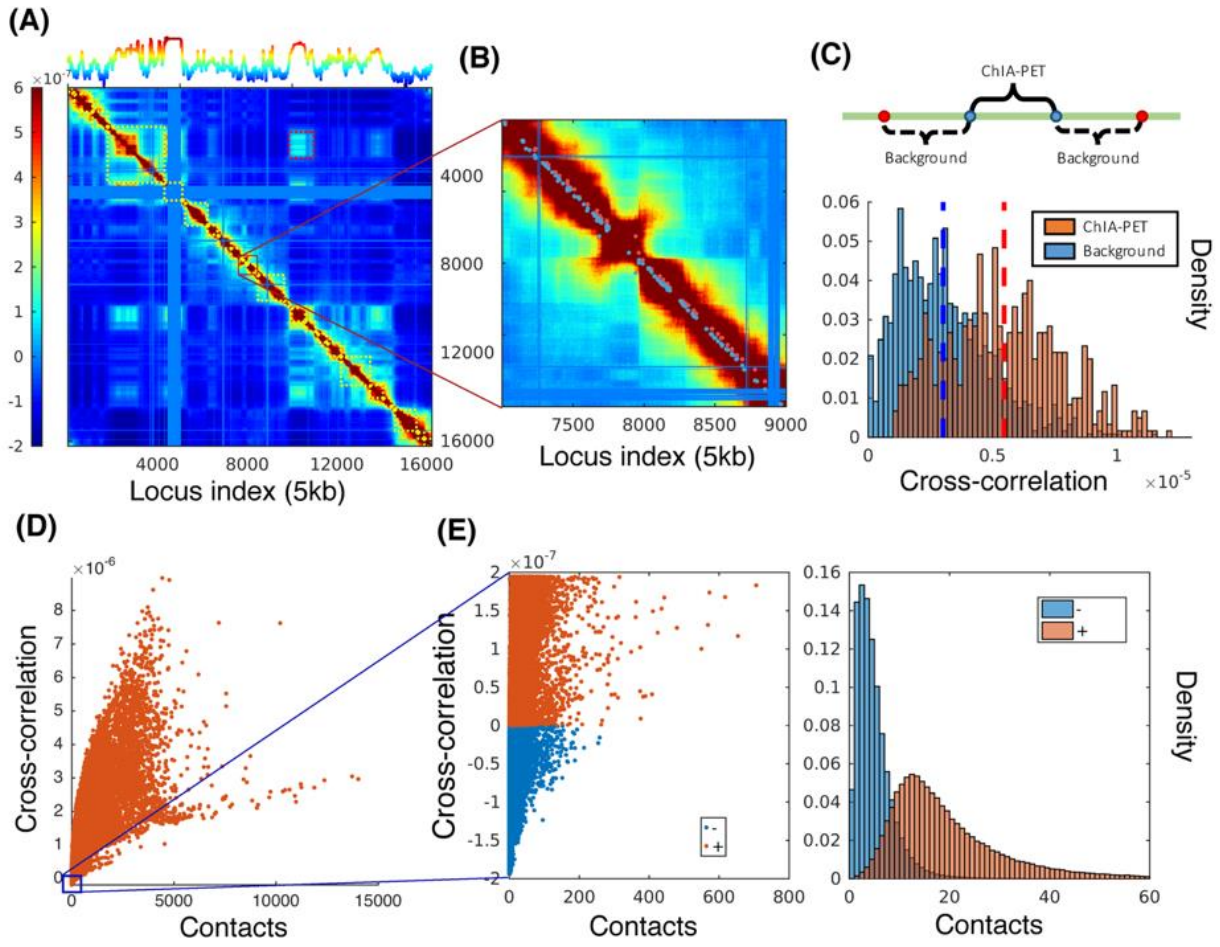

Figure 2.5 Covariance map computed for chromosome 17 and comparison with ChIA-PET data and contacts from Hi-C experiments in GM12878.(A) Covariance matrix computed for chromosome 17, color-coded by the strength of cross-correlation between loci pairs (see the *color bar* on the *left*). The *curve* on the *upper abscissa* shows the average overall off-diagonal elements in each column, which provides a metric of the coupling of individual loci to all others. The *blue bands* correspond to the centromere, where there are no mapped interactions. **(B)** Close-up view of a region along the diagonal. *Red dots* near the diagonal indicate pairs (separated by ~100 kb) identified by ChIA-PET to interact with each other; nearby *blue points* are control/background pairs. **(C)** Stronger cross-correlations of ChIA-PET pairs compared to the background pairs. **(D)** Dependence of cross-correlations on the number of contacts observed in Hi-C experiments. A broad distribution is observed, indicating the effect of the overall network topology (beyond local contacts) on the observed cross-correlations. **(E)** Loci pairs exhibiting anti-correlated (same direction, opposite sense) movements usually have fewer contacts, compared to those exhibiting correlated (same direction, same sense) pairs of the same strength. This figure is adapted from (Sauerwald, et al., 2017).

2.1.2.4 Cross-correlations between loci motions encoded by chromosomal network topology

In general, loci-loci cross-correlations become weaker with increasing distance along the chromosome, and some pairs show anticorrelations (i.e. move in opposite directions; see *color bar* in **Figure 2.5A**). Yet, we can distinguish distal regions that exhibit notable cross-correlations in the spatial movements (off-diagonal *lighter-colored* blocks). The levels of cross-correlations do not necessarily need to scale with the interaction strengths between the correlated loci (or number of contacts detected by Hi-C). On the contrary, a broad range of cross-correlations is observed for a given number of contacts, indicating that the observed correlations are global properties defined by the entire network topology and reflect the collective behavior of the entire structure.

Figure 2.5D displays the computed cross-correlations as a function the number of contacts, showing that some pairs of loci display much stronger correlations revealed by the GNM than others that make more Hi-C contacts. **Figure 2.5E** shows that the anticorrelated pairs of loci (*blue*) usually have fewer contacts than those (*red*) exhibiting positive cross-correlations of the same strength. This analysis thus shows that gene loci pairs which exhibit the same frequency of contacts in Hi-C experiments may have stronger or weaker cross-correlations in their spatial movements depending on the overall topology of the network.

2.1.2.5 Dynamically correlated distal regions exhibit higher co-expression

The GNM covariance map further shows the existence of correlations between the movements of farther apart (>10 Mb) regions. In contrast to the main diagonal, the majority of the off-diagonal space typically shows significantly weaker correlations. Regions in this space with higher than expected covariance values represent dynamically linked windows along the chromosome, which may represent long-range interactions. We call these pairs of windows *cross-correlated distal domains* (CCDDs). To identify CCDDs, we set a threshold for each covariance

matrix equal to the absolute value of the minimum covariance. Treating the remaining adjacent pairs as edges in a graph, we then locate connected components beyond the widest section of the main diagonal and above the threshold that contain more than one bin pair, and find the maximal-area rectangle contained within each connected region of high covariance values. These CCDDs are therefore pairs of regions distant along the chromosome, composed each of highly interconnected loci, which also exhibit relatively high cross-correlations compared to other regions of similar genomic separation. Previous methods for identifying long-range chromatin interactions (Rao, et al., 2014; Sanyal, et al., 2012; Xu, et al., 2016; Yaffe and Tanay, 2011) have focused on locating individual points of interaction within 1–2 Mb apart, while CCDDs tend to be on the order of tens of Mb apart and supported by groups of interacting loci.

The covariance matrix results from the overall coupling of the complete network of loci upon inversion of the Kirchhoff matrix for the entire chromosomes. As such, it permits to capture, or better discriminate, the long-range correlations resulting from the complex topology of loci-loci contacts, as opposed to the raw data on local loci-loci contacts described by Hi-C maps. The covariance data also permit the identification of an appropriate threshold value for defining the significant CCDDs, consistent with the cooperative couplings within the entire structure, including distal correlations. There is no correspondingly clear threshold value for raw Hi-C data, which makes identifying these regions difficult without covariance matrices. Highly distant gene pairs within CCDDs show greater co-expression values than gene pairs outside these regions (p-value $< 10^{-7}$ using the background defined below). For each CCDD, we identified the genes contained within the region and measured the co-expression of each gene pair from distant chromosomal segments. The background gene pairs were gathered from outside the CCDDs but with similar

genomic separation as the CCDD gene pairs. We computed gene expression correlations from 212 experiments (see Section 2.1.4.5).

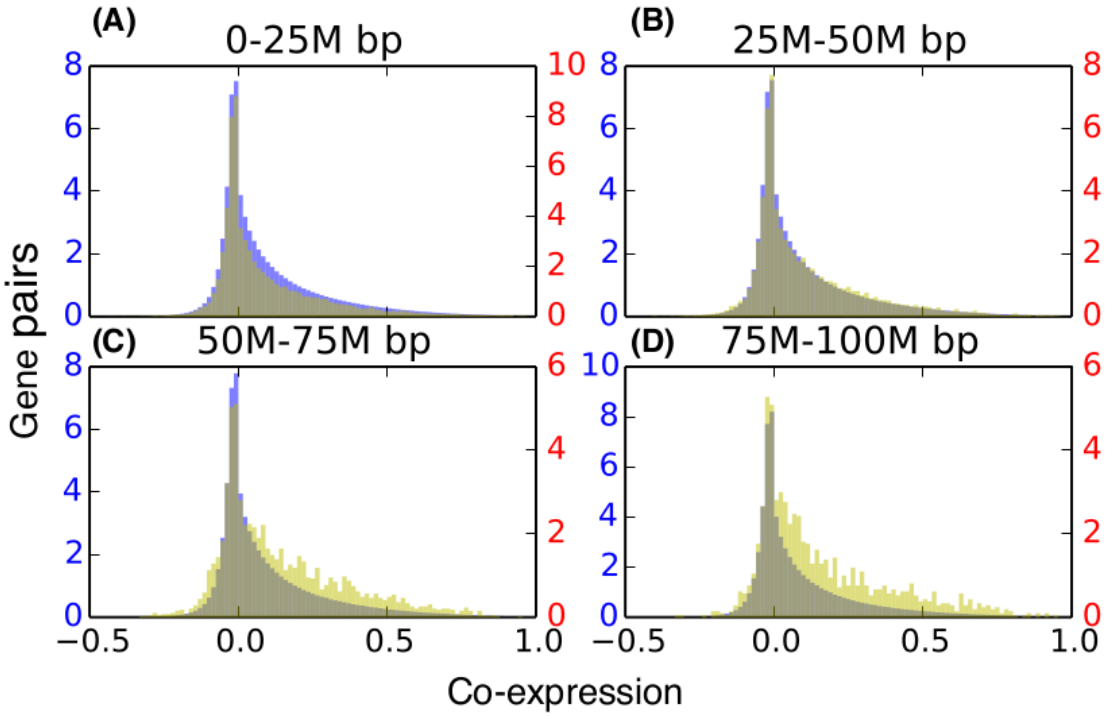

Figure 2.6 Co-expression is significantly enriched in CCDDs. In each histogram, the *yellow* distribution represents gene pairs from CCDDs and the *blue* distribution represents background gene pairs. All are showing the normalized number of gene pairs with a particular Pearson expression correlation for gene pairs within a distance of **(A)** 0-25 Mb, **(B)** 25-50 Mb, **(C)** 50-75 Mb, and **(D)** 75-100 Mb. The more distant pairs (50-100 Mb apart) within the CCDDs show enriched expression correlations as compared to the background pairs. There were not enough gene pairs within CCDDs more than 100Mb apart to draw significant conclusions. This figure is adapted from (Sauerwald, et al., 2017).

As seen in **Figure 2.6**, the CCDDs containing specifically gene pairs that are between 50 and 100 Mb apart are much more highly co-expressed than background gene pairs at the same genomic distance (p-value $< 10^{-19}$; see Section 2.1.4.5 for details). This indicates that the couplings between these genes, as revealed by GNM, may often be biologically important. CCDDs at smaller genomic distance (<50 Mb) exhibit similar co-expression distributions to the

background gene pairs, likely due to the effect of shorter genomic distances including more co-regulated genes within the background. Beyond distances of 100Mb, there are not sufficient gene pairs within CCDDs to draw any meaningful conclusions. Dynamically coupled regions that are very distant sequentially but biologically linked through gene expression are therefore identifiable using the GNM covariance matrix.

Overall, the co-expression enrichment of CCDDs supports the significance of the GNM-predicted distal correlations and showed that the genes that are distant along the linear sequence may be co-regulated through spatial (re)arrangement of corresponding genomic regions that brings the genes into spatial proximity.

2.1.3 Discussion

The GNM is particularly adept at predicting topology-dependent dynamics and identifying long-range correlations - the type of modeling that has been a challenge in chromatin 3D modeling studies. Hi-C matrices, in which each entry represents the frequency of contacts between pairs of genomic loci, can be interpreted as chromosomal contact maps similar to those between residues adopted in the GNM representation of proteins.

There are several differences between the Hi-C and GNM $\mathbf{\Gamma}$ matrices. The first is the size: human chromosomes range from ~50 to 250 million base pairs. When binned at 5kb resolution, this leads to 10,000 – 50,000 bins per chromosome. GNM provides a scalable framework, where the collective dynamics of supramolecular systems represented by 10^4-10^5 nodes (such as the ribosome or viruses) can be efficiently characterized. GNM may therefore be readily used for analyzing intrachromosomal contact maps at high resolution. The second is the precision of the data. Experimental methods for resolving biomolecular structures such as X-ray crystallography,

NMR, and even cryo-electron microscopy yield structural data at a much higher resolution than current genome-wide studies. The Hi-C method is population-based (derived from hundreds of thousands to millions of cells), noisy, incomplete (e.g. unmapped regions). However, the GNM results are usually robust to variations in the precision/resolution of the data on a local scale, and require information on only the overall contact topology rather than detailed spatial coordinates, which supports the utility of Hi-C data and applicability of the GNM. Third, the chromatin is likely to be less 'structured' than the structures at the molecular level, and it is likely to sample an ensemble of conformations that may be cell- or context-dependent. Single-cell Hi-C experiments have indicated cell-cell variability in chromosome structure on a global scale, though the domain organization at the megabase scale is largely conserved (Nagano, et al., 2013). Therefore, structure-based dynamic features may be assessed at best at a probabilistic level.

In this work, we analyzed the chromosome dynamics using an elastic network model, GNM, which has found wide applications in molecular structural biology. Though other models (Chen, et al., 2015; Chen, et al., 2016) have examined genome structure through graph theoretical methods, the inclusion of the complete spectrum of motions in the analysis provides a more realistic picture of chromosomal dynamics in accord with a wealth of experimental data. GNM is a mathematically rigorous approach, based on first physical principles, with intuitive interpretations and well-established theoretical and physical underpinnings. It enables us to evaluate, compare and consolidate a broad range of biologically significant genome-wide properties with the help of a unified model. These properties include the evaluation of the MSFs of loci (using data at 5 kb resolution), the discrimination of short-range regulatory interactions among close-neighboring loci, and the identification of dynamically coupled CCDDs. These respective predictions were shown to satisfactorily compare with data from chromatin accessibility

(DNase-seq and ATAC-seq) and ChIA-PET experiments, and predictions from previous computational methods. The agreement with experiments not only validates the applicability of the GNM, but also provides a new set of independent data, which consolidate those from experiments, especially when the experimental data themselves exhibit some differences (see **Figure 2.2D**). The application to two different cell types also showed that GNM data comply with cell-cell variability. The identification of CCDDs allowed us to locate spatially coupled co-expressed regions of the genome which are vastly distant (over 50 Mb apart) along the chromosomes, and this information cannot be found from gene expression or other experimental data alone.

Due to the fact that the Hi-C experiments suffer from several systematic biases, we chose Vanilla Coverage (VC) normalization to eliminate such biases. The choice of VC normalization was based on the satisfactory correlation between theoretical predictions and chromatin accessibility (see Section 2.1.4.2) (Duan, et al., 2010; Rao, et al., 2014). We further assessed the impact of the simplest bias, GC content, on our results. In order to verify that the agreement with experimental data was not simply due to obvious covariates, we measured the correlation between GC content and both DNase-seq and ATAC-seq. On all chromosomes, the MSFs from GNM exhibited higher correlation with both experimental datasets than GC content. The correlation between GC content and accessibility data averaged to 0.606 and 0.278 for DNase-seq and ATAC-seq, respectively, compared to 0.800 and 0.552 achieved by the GNM-predicted MSFs. Co-expression enrichment of CCDDs was maintained after bias-corrected RNA-seq quantification (data not shown; see Supplementary Figure S17 in (Sauerwald, et al., 2017)), also supporting the significance of the GNM predicted distal correlations. Future efforts may focus on deploying

methods that can remove bias factors from both Hi-C and accessibility data, in order to fully separate the capabilities of GNM from simple covariates.

In general, the evaluation of dynamic features using structure-based models becomes prohibitively expensive with increasing size of the structure, hence the development of coarse-grained models and methods for exploring supramolecular systems dynamics. The chromatin size is well beyond the range that can be tackled efficiently by structure-based methods and realistic force fields. The applicability of the GNM to modeling chromatin dynamics originated in two fundamental features: its scalability and its ability to solve for collective fluctuations and cross-correlations based on network contact topology, exclusively. No knowledge of structural coordinates was needed, nor did we predict structural models - a task that has been undertaken successfully by recent studies (Ay, et al., 2014; Bau and Marti-Renom, 2011; Bau, et al., 2011; Rousseau, et al., 2011; Stevens, et al., 2017; Varoquaux, et al., 2014; Zhang and Wolynes, 2015). We characterized the collective dynamics encoded by the overall chromosomal contact topology, driven by entropy, consistent with the ensemble-based properties of the genome structure. MSFs predicted by the GNM represent ensemble averages over thermal fluctuations (see Section 2.1.4.3), and reflect population-averaged behavior examined in the Hi-C experiments, hence their applicability to population-average based experiments such as ATAC-seq and DNase-seq.

2.1.4 Methods

2.1.4.1 Data preprocessing

Our Hi-C data came from the large, high-resolution Hi-C dataset (GEO accession: GSE63525), pre-processed using VC normalization (Rao, et al., 2014). We used Hi-C data at 5kb resolution unless otherwise noted. DNase-seq data were collected as part of the ENCODE project

(ENCFF000SKV for GM12878 cells, ENCFF740JVK for IMR90 cells) (Consortium, 2004). The ATAC-seq measurements (Buenrostro, et al., 2013) were also obtained for GM12878 and IMR90 cells (GEO accessions GSM1155959 and GSM1418975, respectively). For both experimental datasets, called peaks were binned to the same resolution as the Hi-C data by adding all peak values within each bin. The binned data were then smoothed using a moving average with a window size of 200 kb. The long-range interactions from ChIA-PET were from ENCODE (ENCFF002EMO) (Heidari, et al., 2014). We used a two-sample t-test assuming unequal variances to quantify the difference between the covariance distributions of ChIA-PET and background interactions.

2.1.4.2 Hi-C Data Normalization

We tested three types of normalization methods applied to the Hi-C contact map: VC normalization (referred to as VCnorm), square-root VC normalization (referred to as sqrtVC) (Rao, et al., 2014) and Knight-Ruiz normalization (referred to as KRnorm) (Knight and Ruiz, 2013). All three methods aim to eliminate the so-called "one-dimension bias" (Rao, et al., 2014). We found that the GNM performed best on Hi-C maps normalized by VCnorm when benchmarked against experimental data (**Figure 2.7**). Not only are the correlations with the chromatin accessibility lower, but also the square-fluctuations become increasingly flatter upon inclusion of a higher number of modes in the calculations when KRnorm or sqrtVC are applied on the contact map. In the extreme case, when all the modes are used, the square fluctuations become almost completely flat along the chromosome using KRnorm. This is because KRnorm ensures that every row and column sum up to 1. As a consequence, all loci become almost equally constrained and the differences in their square fluctuations are suppressed.

In addition, computations with the three normalization methods were repeated at different resolutions, and VCnorm yielded the most robust agreement between theoretically predicted MSFs

and experimentally observed accessibilities across all resolutions. Both KRnorm and sqrtVC showed poor correlations at high resolution (5kb) (**Figure 2.7**). Furthermore, VCnorm showed the expected improvement in correlation using increasing number of modes included in the analysis, while KRnorm or sqrtVC led to inconsistent results, even at 50kb resolution (**Figure 2.7**). Due to the better performance across resolutions and numbers of modes, shown by the agreement with experimental data, we chose VC normalized contact maps to perform further analyses.

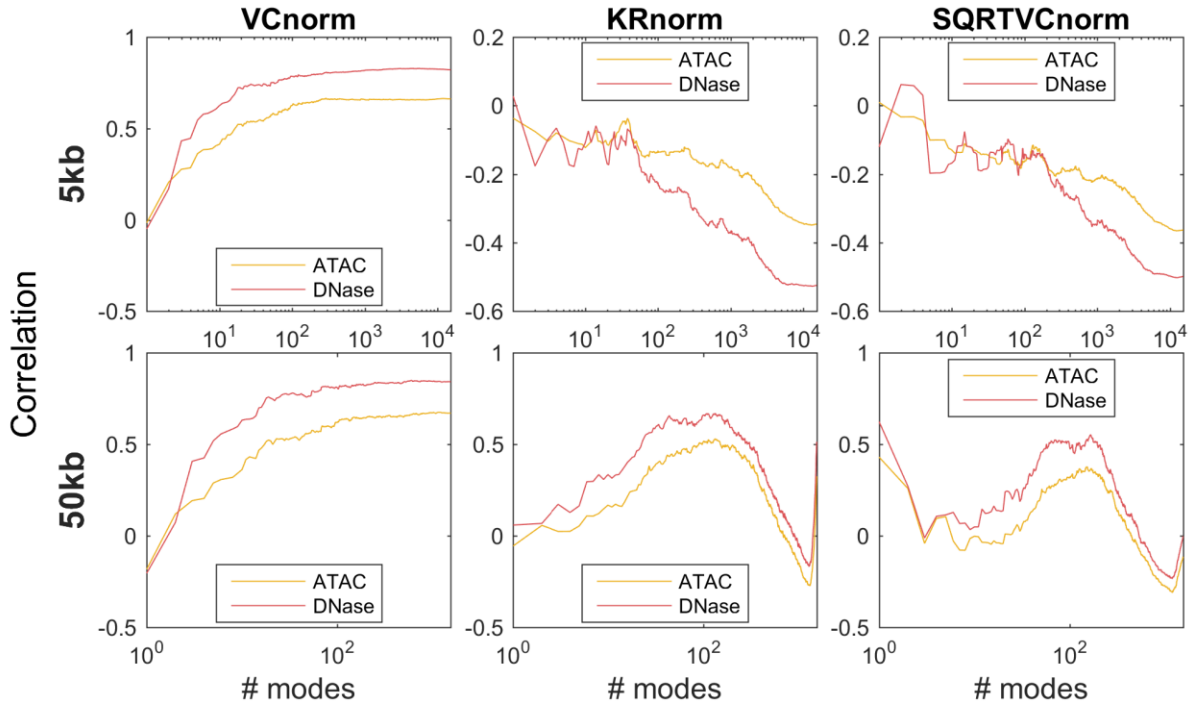

Figure 2.7 The scanning of correlations between chromatin accessibility and square fluctuations calculated as a function of the number of modes included in the GNM analysis. The rows compare the correlations at different resolutions, and the columns compare those computed from three different normalization methods. Note the poor performance of KRnorm and SQRTVCnorm, especially in the case of high resolution data (5kb). This figure is adapted from (Sauerwald, et al., 2017).

2.1.4.3 Extension of the GNM to modeling chromatin dynamics

The GNM was first proposed in 1997 for analyzing protein dynamics (Bahar, et al., 1997). Similar to the ANM, the GNM describes the structure as a network of nodes connected by elastic springs. The two models differ in that the GNM uses exclusively the inter-residue contact topology and ignores the 3D coordinates of the residues. Instead of using a $3n \times 3n$ Hessian, n being the number of nodes in the network, the network topology in the GNM is defined by an $n \times n$ Kirchhoff matrix $\mathbf{\Gamma}$, whose elements are

$$\mathbf{\Gamma}_{ij} = \begin{cases} -\gamma_{ij} & d_{ij} \le d_0 \text{ and } i \ne j \\ 0 & d_{ij} > d_0 \text{ and } i \ne j \\ -\sum_{j,j\ne i} \gamma_{ij} & i = j \end{cases}. \tag{2.1}$$

Here γ_{ij} represents the strength or stiffness of interaction between beads i and j (or the force constant associated with the spring that connects them), d_{ij} is their distance separation in the 3D structure, and d_0 is the cutoff distance for making contacts (or for being connected by a spring). The GNM potential is defined as

$$V_{GNM} = \frac{1}{2}\sum_{i,j} \gamma_{ij} \left(\boldsymbol{d}_{ij} - \boldsymbol{d}_{ij}^0\right)^2 = \frac{1}{2}\sum_{i,j} \gamma_{ij} \left(\Delta\boldsymbol{r}_i - \Delta\boldsymbol{r}_j\right)^2, \tag{2.2}$$

where $\Delta\boldsymbol{d}_{ij}$ and $\Delta\boldsymbol{d}_{ij}^0$ represent the distance vectors, as opposed to scalars in equation 1.1, between node i and j. This difference allows the GNM to take account of the energy changes incurred during internal rotational motions which are neglected in the ANM. $\Delta\boldsymbol{r}_i$ and $\Delta\boldsymbol{r}_j$ are the displacements of node i and j from their equilibrium positions in 3D. The GNM potential can be also represented in a matrix form using the arrays of x, y, z components of $\Delta\boldsymbol{r}$'s, i.e. $\Delta\boldsymbol{x} = [\Delta x_1, \Delta x_2, \ldots, \Delta x_n]$:

$$V_{GNM} = \frac{1}{2}(\Delta\boldsymbol{x}^T\mathbf{\Gamma}\Delta\boldsymbol{x} + \Delta\boldsymbol{y}^T\mathbf{\Gamma}\Delta\boldsymbol{y} + \Delta\boldsymbol{z}^T\mathbf{\Gamma}\Delta\boldsymbol{z}). \tag{2.3}$$

Assuming the fluctuations of the network nodes to be *isotropic*, their probability distribution can be expressed using a Boltzmann factor as

$$p(\Delta \boldsymbol{r}) = p(\Delta \boldsymbol{x})p(\Delta \boldsymbol{y})p(\Delta \boldsymbol{z}) \sim \exp\left\{-\frac{3}{2k_B T}\Delta \boldsymbol{x}^T \boldsymbol{\Gamma} \Delta \boldsymbol{x}\right\}, \tag{2.4}$$

where k_B and T are the Boltzmann constant and absolute temperature. Obviously, $p(\Delta \boldsymbol{r})$ is a multivariate Gaussian distribution which means that the GNM assumes that the network nodes normally fluctuate around their equilibrium positions.

In the application to proteins, the beads represent individual residues, their positions are identified with those of the α-carbons, and a uniform force-constant $\gamma_{ij} = \gamma$ is adopted for all pairs ($1 \leq i, j \leq n$), with a cutoff distance of $d_0 \sim 7$Å. In the extension to the chromatin, we redefine the network nodes and springs such that beads represent genomic loci consistent with the resolution of the Hi-C data. We set γ_{ij} equal to γz_{ij} where z_{ij} is the Hi-C contact counts reported for the pair of genomic bins (Levy-Leduc, et al., 2014) *i* and *j* after normalization by vanilla coverage (VC) method (Rao, et al., 2014), and γ is taken as unity. The element $\boldsymbol{\Gamma}_{ij}$ is thus taken to be directly proportional to the actual number of physical contacts between the loci *i* and *j*, which permits us to directly incorporate the strength of interactions in the network model. The parameter γ uniformly scales all elements, physically representing the strength (or spring constant) of individual contacts.

2.1.4.4 Prediction of the dynamics of genomic loci using the GNM

The covariance of the multivariate Gaussian distribution, $\mathbf{C}_{ij}$, between the spatial displacements of loci *i* and *j* can be obtained from the pseudoinverse of $\boldsymbol{\Gamma}$, as

$$\mathbf{C}_{ij} = \langle \Delta \boldsymbol{r}_i \,.\, \Delta \boldsymbol{r}_j \rangle \sim [\boldsymbol{\Gamma}^{-1}]_{ij} = \sum_{k=1}^{n-1} \frac{1}{\lambda_k} \left[\boldsymbol{v}_k \boldsymbol{v}_k^{\mathrm{T}}\right]_{ij}, \tag{2.5}$$

where the summation is performed over all modes of motion intrinsically accessible to the network, obtained by eigenvalue decomposition of $\mathbf{\Gamma}$. The respective frequencies and shapes of these modes are given by the n - 1 non-zero eigenvalues (λ_k) and corresponding eigenvectors ($\boldsymbol{v}_k$). The eigenvector $\boldsymbol{v}_k$ represents the normalized displacements of the n loci along the k^{th} mode axis, and $1/\lambda_k$ rescales the amplitude of the motion along this mode. In the Hi-C map there are regions where no cross-linked DNA fragments can be mapped. These unmapped regions are isolated from the system, and their existence may lead to multiple zero-eigenvalue modes. These unmapped regions are not constrained by other loci, so they may cause large fluctuations that obscure the signal from other regions. These extra zero-eigenvalue modes and unphysically large fluctuations were removed by discarding the unmapped regions. Note that the removal of the unmapped regions will not cause disconnections because the chromosomes are highly compact, so the loci next to the unmapped regions remained connected to the loci located at the other end of the region.

The i^{th} diagonal element of $\mathbf{C}$, $\langle \Delta \boldsymbol{r}_i^2 \rangle$, is the predicted mean-square fluctuation (MSF) of the i^{th} locus under physiological conditions, which is inversely proportional to the elastic spring constant γ. The MSF profiles thus provide a measure of the relative size of motions of the different gene loci (irrespective of γ), exclusively defined by the particular loci-loci contact topology. They represent ensemble averages over all accessible motions to a given locus.

Cross-correlations in the GNM are easily obtained from the $n \times n$ covariance matrix as

$$\tilde{\mathbf{C}}_{ij} = \frac{\mathbf{C}_{ij}}{\sqrt{\mathbf{C}_{ii}\mathbf{C}_{jj}}}, \tag{2.6}$$

so that the covariances are normalized by the MSFs of corresponding loci. This is particularly helpful in mitigating the cases where some loci experience extremely high or low fluctuations that

dominate the covariances with other loci. The described calculations are summarized and depicted schematically in **Figure 2.8**.

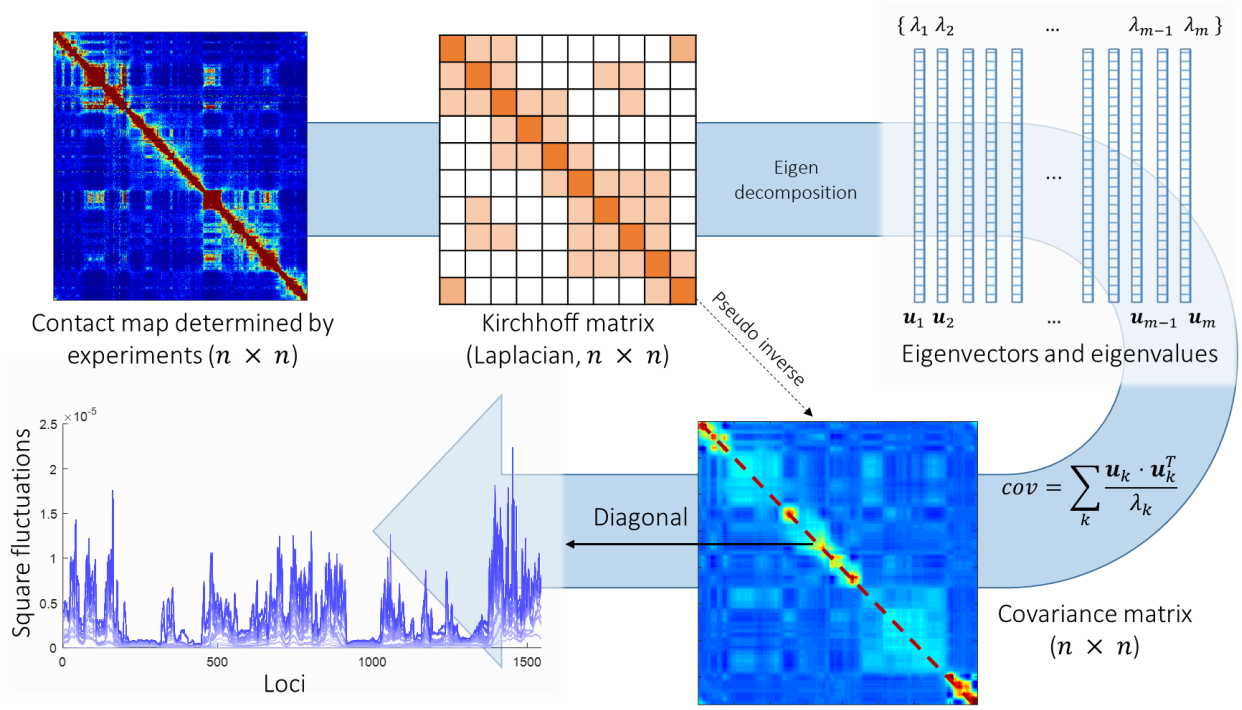

Figure 2.8 Schematic description of the GNM methodology applied to Hi-C data. The inter-loci contact data represented by the Hi-C map (*upper left*, for *n* genomic bins (loci)) is used to construct the GNM Kirchhoff matrix, **Γ** (*top*, *middle*). Eigenvalue decomposition of **Γ** yields a series of eigenmodes which are used for computing the covariance matrix (*lower*, *right*), the diagonal elements of which reflect the mobility profile of the loci (*bottom*, *left*), and the off-diagonal elements provide information on locus-locus spatial cross-correlations. $\boldsymbol{u_k}$, *k*th eigenvector; $\boldsymbol{\lambda_k}$, *k*th eigenvalue; *m*, number of nonzero modes, starting from the lowest-frequency mode, included in the GNM analysis ($m \leq n$-1). In the present application to the chromosomes, *n* varies in the range $10{,}248 \leq n \leq 49{,}850$, the lower and upper limits corresponding respectively to the respective chromosomes 22 and 1. This figure is adapted from (Sauerwald, et al., 2017).

2.1.4.5 Evaluation of co-expression levels

In order to calculate co-expression values for genes in this cell type, we downloaded every publicly available RNA-seq experiment on GM12878 cells from the Sequence Read Archive

(Kodama, et al., 2012), which gave 212 data sets. These raw read data were quantified using Salmon (Patro, et al., 2017), resulting in 212 transcripts per kilobase million (TPM) values for every gene. Quantification was performed with and without bias correction, with qualitatively similar results. Co-expression was then measured as the Pearson correlation of the two vectors of TPM values for a given gene pair.

2.1.5 Acknowledgment

The present chapter was part of the publication (Sauerwald, et al., 2017). Natalie Sauerwald and myself equally contributed to this work, and shared the first authorship. Among the work presented in this section, N.S. identified the CCDDs and performed the co-expression analysis. I developed and implemented the extension of the GNM framework for adapting it to modeling chromatin dynamics and compared GNM predications with chromatin accessibility and ChIA-PET interactions. Drs. Ivet Bahar and Carl Kingsford supervised the project.

2.2 Identification of Hierarchical Chromosomal Domains

2.2.1 Introduction

Spatial organization of the genome is largely compartmental and hierarchical. Instead of mixing with each other, each chromosome is organized by itself in different chromosomal territories (Cremer and Cremer, 2010; Meaburn and Misteli, 2007; Mirny, 2011). Inside each chromosome, the loci that have different interaction patterns can be generally categorized into transcriptionally active and inactive regions, i.e. A and B compartments, respectively (Lieberman-Aiden, et al., 2009; Rao, et al., 2014). Each compartment is typically ~100Mb in size and can be further segregated into smaller parts called topologically associating domains (TADs), which are nested substructures and varied in size (Dixon, et al., 2012; Rao, et al., 2014; Rowley and Corces, 2018). Based on the strengths of inter- and intra-domain interactions, the boundaries of these domains show different levels of clarity. Architectural proteins are highly enriched at domain boundaries, especially at the strongly distinguishable boundaries (Gomez-Diaz and Corces, 2014; Van Bortle, et al., 2014).

Significant progress has been made towards computationally identifying these domains (An, et al., 2019; Durand, et al., 2016; Filippova, et al., 2014; Rao, et al., 2014; Serra, et al., 2017; Weinreb and Raphael, 2016; Yan, et al., 2017; Zhan, et al., 2017); however, many computational methods focus on identifying the domains at a given scale or a given level of resolution, such that a comprehensive characterization of the hierarchical organization of the multiscale genome (simultaneously revealing TADs, A/B compartments and other levels) has been elusive.

The intrinsic dynamics identified by the GNM includes a spectrum of independent spatial movements, i.e. normal modes. Each mode k is characterized by two properties: its vibrational

frequency and "shape", described by the respective k^{th} eigenvalue and eigenvector of the Kirchoff matrix $\mathbf{\Gamma}$ representing the network topology. The frequencies are indicative of the relative time scales of the modes, the low (high) frequency modes usually corresponding to global (local) motions that involve large (small) parts of the structure. Therefore, the frequencies in the GNM can be regarded as a "resolution parameter" that defines the granularity of the structure whose dynamics we are examining. Notably, the GNM shares a very similar mathematical formulation with spectral clustering (see Appendix A), and so the GNM modes can be readily used to identify chromosomal domains. However, despite the physical insights provided by the GNM, the first *k* modes merely yield *k* clusters and do not necessarily guarantee a hierarchical or nested division as *k* increases, which may pose a challenge for obtaining a hierarchical view based on clusters identified using different *k*.

To address this problem, we developed a network analysis framework, termed Hierarchical community Decoding Framework (HiDeF). HiDeF enables us to determine the hierarchical relationships among the structural domains identified by the GNM modes. (see Section 2.2.4.4). Taking a Hi-C contact map as input, chromosomal domains are identified at many resolutions, after which containment relationships are systematically inferred for pairs of domains. The result is a directed acyclic graph (DAG) representing the inferred hierarchical domains, in which vertices at increasing distances from the root represent domains of increasing granularity. This framework has parameters which allow for flexible control of model complexity: the output hierarchy can be simple, prioritizing only the strongest patterns in data, or complex, retaining increasing numbers of auxiliary patterns that arise during the resolution sweep.

We applied this framework, described in detail in the Methods subsection 2.2.4.4, to the GNM analysis of Hi-C chromatin interaction data. The hierarchical representation of these overall

network structure uncovered significant structural properties relevant to functional entities, and yielded novel insights into the spatial organization of genomes beyond that often described in terms of compartments and/or TADs.

2.2.2 Results

2.2.2.1 Different types of (sub)compartments show differentiated levels of mobility and accessibility

A/B compartments were first discovered in an early Hi-C study of the human chromosomes (Lieberman-Aiden, et al., 2009). It was found that the whole genome can be split into two spatial compartments that correspond to the open and closed chromatin (labeled as compartments A and B, respectively) based on the distribution of loci along the first principle axis/component deduced from a PCA analysis of processed Hi-C contact matrices. Regions within each type of the compartment tend to interact with others of the same type and much less frequently with others of the different type. Later, with advances in experimental techniques and availability of higher resolution Hi-C data, six subtypes (A1, A2, B1, B2, B3, B4) of compartments were discovered based on their long-range interaction patterns, both within and between chromosomes (Rao, et al., 2014).

Here we examined the chromatin mobility (predicted by the GNM) and accessibility (measured by ATAC- or DNase-seq experiments) of the genomic regions labeled as open (compartment A) and closed chromatin (compartment B). We confirm that the regions in compartment A experience significantly greater mobility and accessibility than those in compartment B (p-value < 0.01, see **Figure 2.9**).

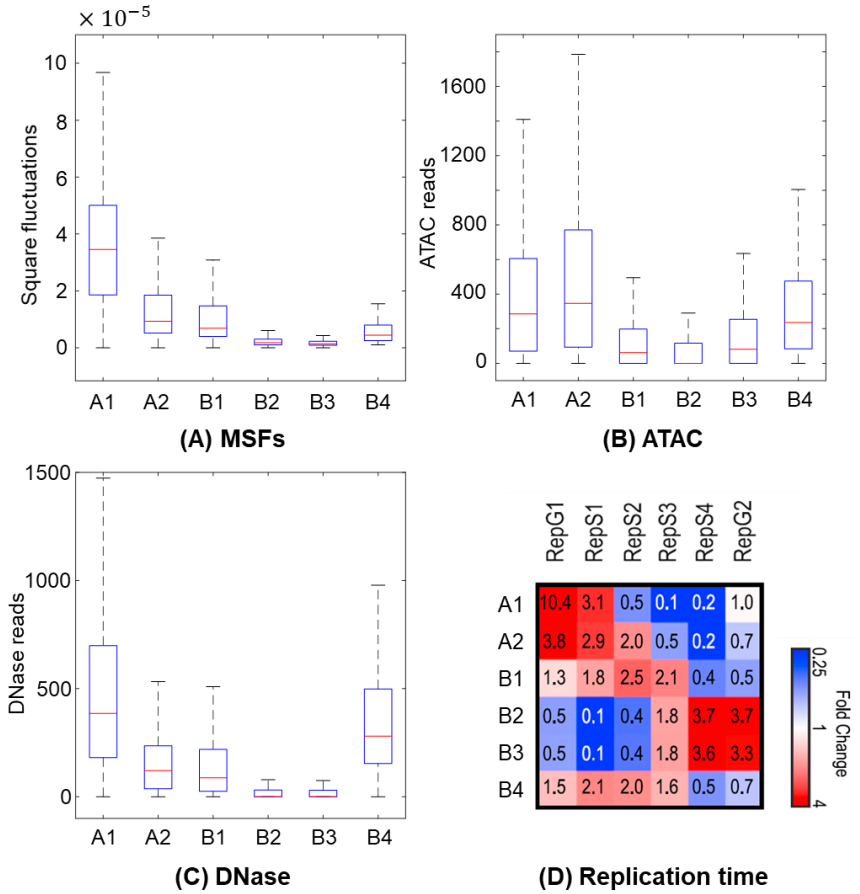

Figure 2.9 Chromatin mobility and accessibility of regions belonging to different subcompartments. The *box plots* show the distributions of **(A)** GNM-predicted MSFs, **(B)** ATAC-seq measured accessibility, **(C)** DNase-seq measured accessibility of loci grouped based on six compartmental subtypes. The mean value, standard deviation, and min-max range of each distribution are indicated by the *red horizontal line*, the *blue box*, and the *black bars*, respectively. Results were obtained for GM12878 cell line. **(D)** Replication times of different compartments. The heatmap shows the enrichment of regions belonging to different types of compartments being replicated during G1, S1, S2, S3, S4, and G2 phases of the cell cycle. This panel is adapted from (Rao, et al., 2014).

A closer inspection into the subtypes of compartments revealed heterogeneous chromatin mobility and accessibility among regions of different subtypes within the same compartment type. Specifically, compartment A2 seems to be more rigid (compactly folded) than A1, and consequently experiences loci mobility (supported by GNM-predicted MSFs; **Figure 2.9A**) and

accessibility (supported by DNase-seq but not ATAC-seq; **Figure 2.9B** and **C**) almost as low as those of compartment B. This is also supported by the finding that "A2 is more strongly associated with the presence of H3K9me3 than A1" (Rao, et al., 2014), as H3K9me3 is often associated with heterochromatin. Among the subtypes of compartment B, B2 and B3 were shown to be the most rigid subtypes, consistent with the finding that they are enriched at nucleolus-associated domains (NADs) or lamina-associated domains (LADs). B1 and B4, on the other hand, showed relatively enhanced mobilities and accessibilities with respect to other B subtypes. Overall, the mobility and accessibility profiles of compartment subtypes coincide with replication times of the regions in that the more mobile (accessible) regions tend to replicate sooner in the cell cycle and *vice versa* (**Figure 2.9D**), suggesting that the spatial mobility/accessibility is an important property that affects the regulation of gene transcriptional activities.

2.2.2.2 Domains identified by GNM at different granularities correlate with known structural features

Compartments are multi-megabase-sized regions in the genome characterized by known genomic features such as gene presence, levels of gene expression, chromatin accessibility, and histone markers (Lieberman-Aiden, et al., 2009; Rao, et al., 2014). As mentioned above, Hi-C experiments have revealed two broad classes of compartments: "A" compartments are generally associated with active (open) chromatin, containing more genes, fewer repressive histone markers, and more highly expressed genes; and "B" compartments are associated with less accessible DNA (closed chromatin), have sparser genes, and exhibit higher occurrence of repressive histone marks. TADs (Dixon, et al., 2012) are finer resolution groupings of the chromatin distinguished by denser self-interactions and associated with characteristic patterns of histone markers and CCCTC-binding factor (CTCF) binding sites near their boundaries. The multiscale nature of GNM

spectrum allows for exploring the hierarchical levels of organization within the system. It is of interest to examine to what extent these two levels can be detected by GNM analysis of chromatin structural hierarchy.

As presented above, the GNM LF modes reflect the global dynamics of the 3D structure, and increasingly more localized motions are represented by higher frequency modes. We identified domains from subsets of GNM modes that group regions of similar dynamics (see Section 2.2.4.3). In order to verify whether these GNM-predicted domains correspond to TADs at various resolutions, we used the TAD caller Armatus (Filippova, et al., 2014), varying its "γ" parameter that controls resolution. We refer to this latter parameter as the Armatus γ_k, to distinguish it from the force constant in the GNM. We measured the agreement between GNM domains and TADs using the variation of information (VI) distance, which computes the agreement between two partitions, and where a lower value indicates greater agreement (Meilă, 2003). For more information on the VI metric, see Section 2.2.4.2. For each choice k of number of modes adopted for GNM partitioning of the chromatin structure, the Armatus γ_k that minimizes the VI distance between the GNM domains and the Armatus domains was selected. This resulted in a mean VI value of 1.251 for optimal parameters, which is significantly lower than the VI distance of 1.946 obtained when the GNM domains were randomly re-ordered along the chromosome and compared back to the original TADs (empirical p-value < 0.01 for all chromosomes). **Figure 2.10A**, *left panel* shows the comparison for each chromosome between (i) the VI values obtained for the optimally matched TAD boundaries with the GNM domains and (ii) the distribution of VI values obtained upon random shuffling of domains. As the number of included GNM modes increases, γ_k monotonically increases as well, showing that the number of GNM modes is a good proxy for the scale of chromatin structures sought.

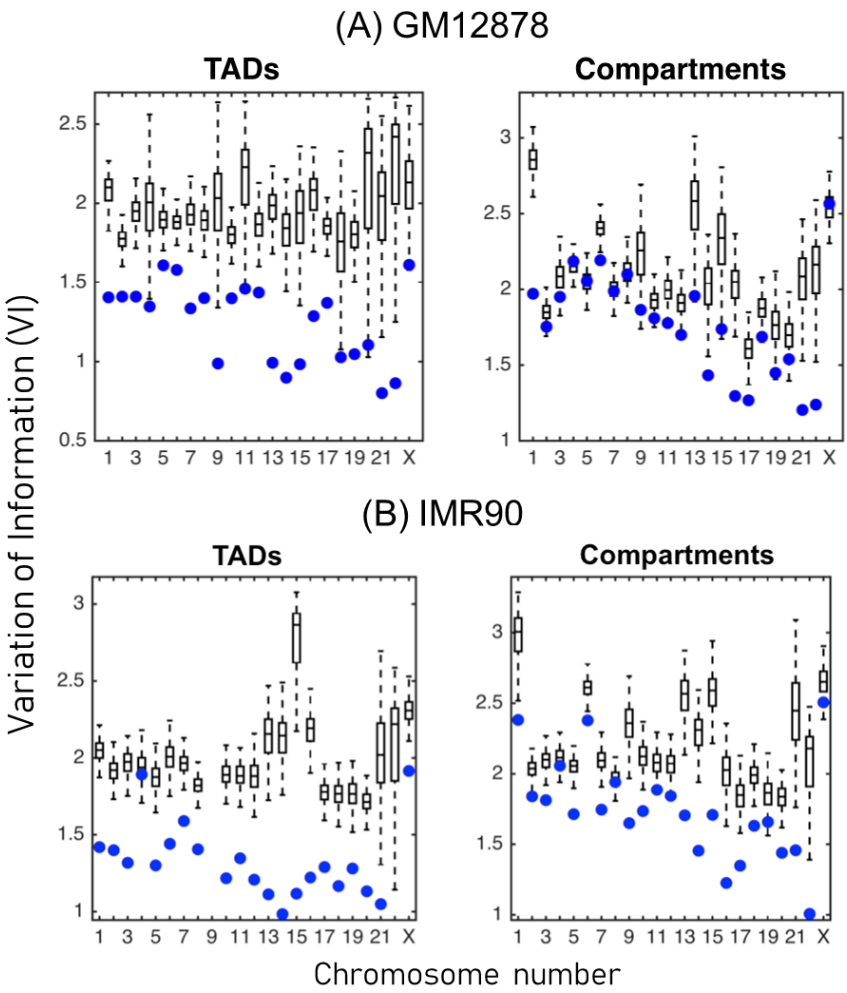

Figure 2.10 Comparison of GNM domains with TADs and Compartments. Results were obtained for **(A)** GM12878 and **(B)** IMR90. Variation of information (VI) measures for comparing GNM domains with TADs (*left panels*) and compartments (*right panels*). Lower VI indicates greater agreement. *Box plots* show the distribution of VI values obtained by randomly shuffling GNM domains and comparing to original TAD and compartment boundaries. *Blue dots* represent the VI value of the true GNM domains with TADs and compartments, respectively. This figure is adapted from (Sauerwald, et al., 2017).

Furthermore, the GNM predicts large-scale global motions using a relatively low number of modes, so we compared these regions discerned in LF modes to larger-scale compartments. We found that the first 5-20 non-zero modes correspond fairly well to compartments. For each

chromosome, we selected the number of modes that produced the smallest VI distance between Lieberman-Aiden compartments and GNM domains. This yielded a mean optimal VI distance of 1.771 (using an average of 13 modes; **Figure 2.11**). This is significantly lower than the mean optimal VI distance of 2.088 when the locations of Lieberman-Aiden compartments are randomly shuffled along the chromosome, though the difference is only statistically significant for 16 of the 23 chromosomes, with *p*-value equal to 0.05. The comparisons of GNM domains with compartments for each chromosome in GM12878 cells can be seen in **Figure 2.10A**, *right panel*. The same calculations were performed on IMR90 cells, with qualitatively similar results. For the comparisons with randomly shuffled domains on IMR90 cells, the results for only 1 chromosome for TADs and 3 chromosomes for compartments were statistically insignificant (**Figure 2.10B**).

Figure 2.11 further shows the GNM domains found using the number of modes that minimizes the VI with compartments or TADs at a lower (50kb) and higher (10kb) resolution. Interestingly, despite having five times more modes in the higher resolution, in both cases it takes ~35 and ~12 GNM modes to capture TAD- and compartment-like structures, respectively. The ability of the GNM to recapitulate both TADs and compartments - two organizational levels of wildly different scales - shows the flexibility and generality of the GNM approach.

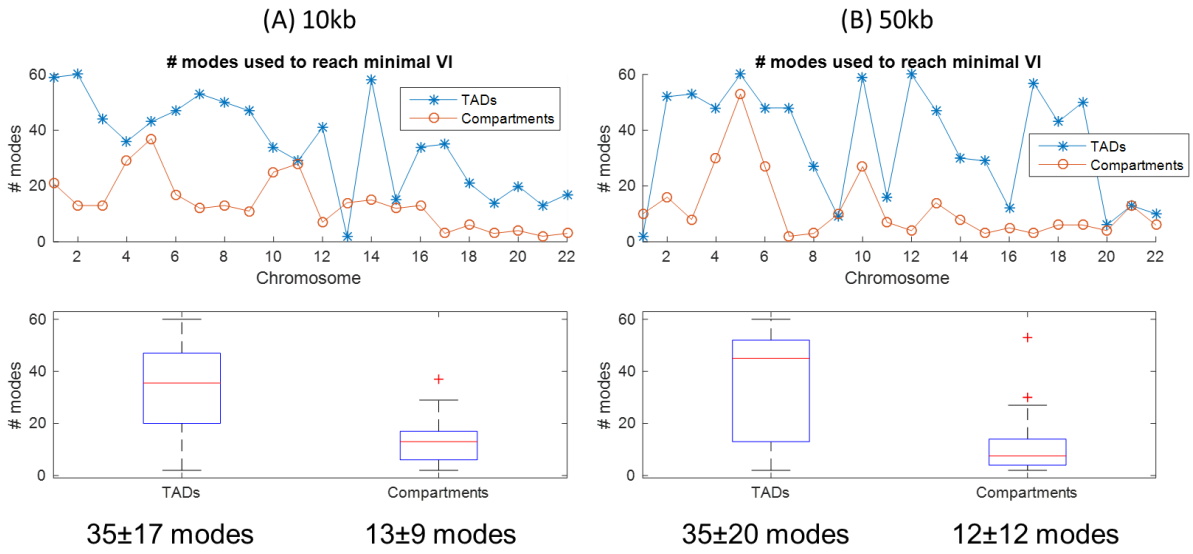

Figure 2.11 The number of modes used to find GNM domains that minimizes the VI with compartments or TADs. Results were obtained using **(A)** 10kb and **(B)** 50kb resolution Hi-C data for GM12878. *Upper panels* show the results for the individual chromosomes. *Lower panels* show the averages and distributions of the results when all chromosomes are considered together. The fact that panels **A** and **B** show similar results suggests that the number of GNM modes used for finding compartments or TADs is insensitive to the resolution.

2.2.2.3 Hierarchical spatial organization of the mouse genome

In the previous sections, we have focused on analyzing the chromosomal domains at a given resolution. Now, equipped with the methodology that allows us to identify domains at multiple scales, and the aforementioned framework, HiDeF, that can infer hierarchical structure from multiresolution clusterings, we proceed to obtaining a hierarchical view of the spatial organization of the genome using the GNM modes (see Section 2.2.4.3). Hierarchical domains identified by HiDeF (see Section 2.2.4.4 and Appendix B) are illustrated in **Figure 2.12** for mouse embryonic stem cells (mESC) chromosome 5, as an example. The corresponding domains (outlined in the *lower triangle* of the matrix in **panel A**) are found to be organized into a hierarchy of 15 levels/depths (see dendrogram in **panel B**). On the first level (the "zeroth" level merges the

entire chromosome into one domain), the densely packed tail (*top-right* part of the matrix, in **panel A**) is reasonably considered as one domain, and the loosely packed head-to-body (*bottom-left* to *middle* part of the matrix), as two additional domains (which might be merged into one, arguably). These gigantic domains are then further divided and separated into smaller and finer domains as the depth increases. Notably, GNM coupled with HiDeF provides a much clearer and more interpretable view of the nested domains than those originally identified by GNM modes alone.

Chromosome banding is a technique used for producing a visible karyotype by staining condensed chromosomes (Speicher and Carter, 2005). In the most used Giemsa (G)-staining technique, the heterochromatin regions (closed chromatin) stain is darker, in contrast to euchromatin (open chromatin) which incorporates less stain and appears as light bands. The resulting chromosome bands or cytobands divide the chromosome into different parts, in a hierarchical fashion, as the nomenclature of the bands involves several levels. Typically, the first level defines the chromosome arms. For example, each human chromosome has two arms, i.e. a short arm (denoted by "p") and a long arm ("q"), separated by the centromere. In the case of the mouse genome, there is only one arm ("q") for each chromosome. Then, the arms can be further divided into regions denoted by capital letters and then numbers to capture more refined banding patterns (e.g. 5qB3.1 refers to a stained region on the long arm of chromosome 5, major G-band B). Even though such division by the chromosome banding can be crude (~10Mb) as compared to the size of compartments and TADs, and the experiment is designed to mark the parts of a metaphase chromosome (whereas Hi-C experiments typically operate on interphase chromosomes), chromosome bands still provide good experimental indications of where the open and closed chromatin regions are and how they are hierarchically organized (see *upper diagonal* parts of **Figure 2.12** panel **A** and the dendrogram in panel **C**), which can be conveniently compared

with our computationally derived model. Such a comparison is illustrated in the panels **D** and **E** of **Figure 2.12** and in **Figure 2.13.**

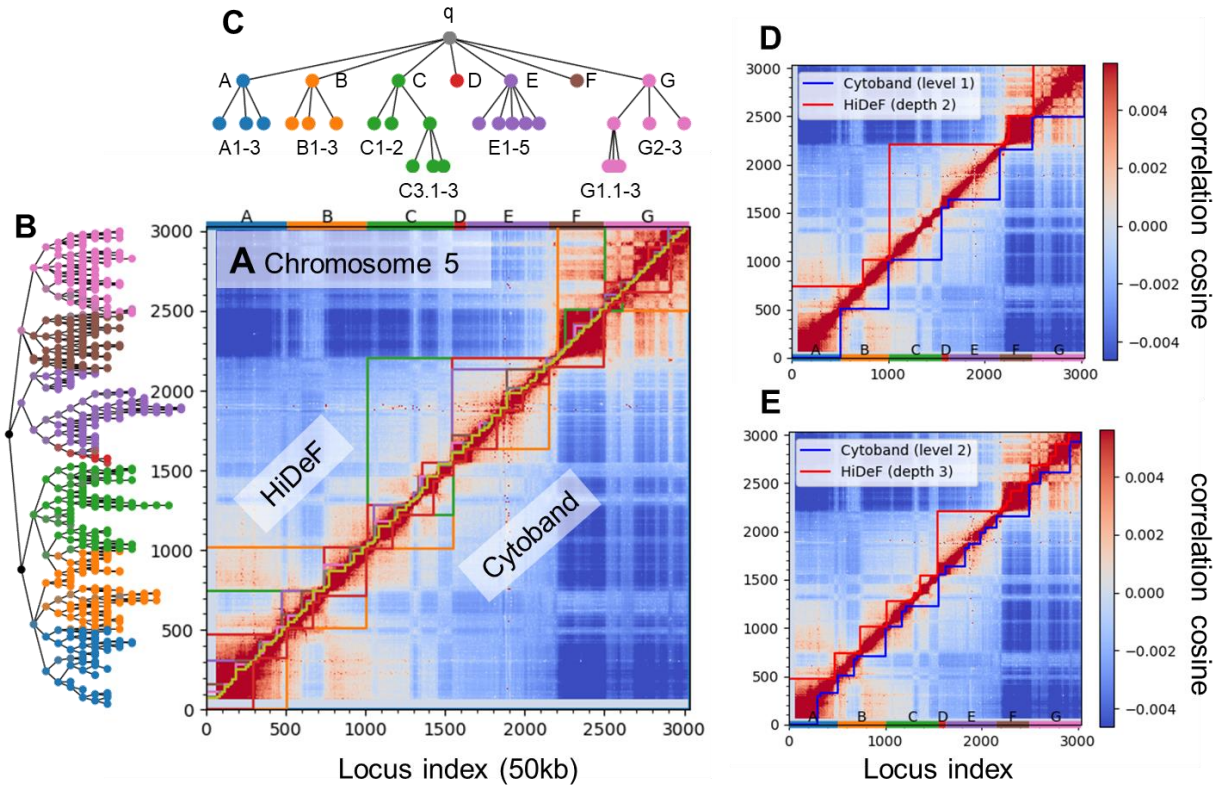

Figure 2.12 Hierarchical organization of mESC chromosome 5 structure. (A) Cross-correlation map of loci movements. The colored lines in the *upper* and *lower triangle* of the map outline the structural domains defined by chromosome bands and GNM modes, respectively. The color bar on *top* indicates the ranges of the cytogenetic bands of mouse chromosome 5 (first level, e.g. qA). **(B)** Domain hierarchy determined by HiDeF based on GNM-identified domains. Each *node* represents a GNM domain identified by a certain number of modes and is color-coded by its most overlapping chromosome band. *Dark shade* of the color indicates low overlap (measured by containment index, see equation 2.10), and in the extreme cases, nodes/domains cannot be assigned to a unique chromosome band is colored *black*. **(C)** The implied domain hierarchy by chromosome bands. Each *node* represents a band/domain colored consistently with the *color bar* in panel A. **(D)** The first level of chromosome bands (e.g. qA) compared with the second level of HiDeF hierarchy. **(E)** The same comparison for the second level of chromosome bands and third level of HiDeF hierarchy.

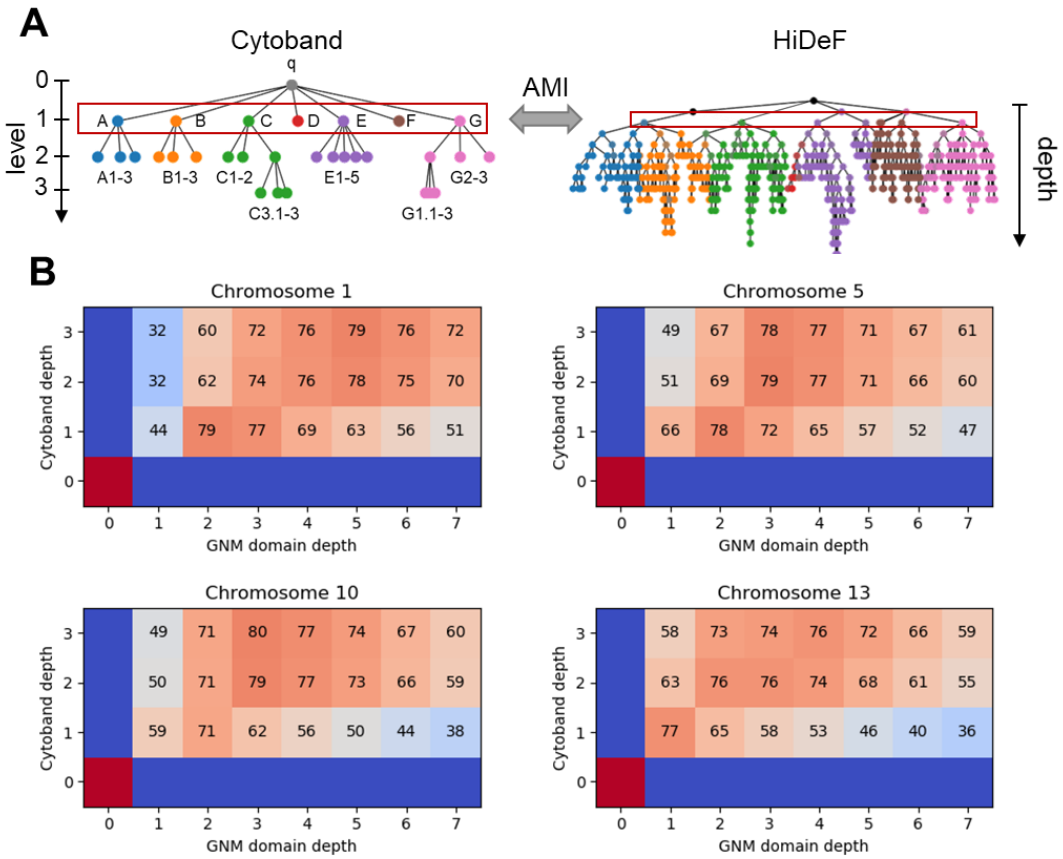

Figure 2.13 Mutual information between chromosome bands and HiDeF/GNM domains defined at different depths. (A) Schematic showing how chromosome bands of different depths are compared with the HiDeF domains resolved at different depths. The "depth" starts at level 0 for both cytoband and HiDeF hierarchies (see the scale on the *left*), which represents the chromosome as an intact domain. The domains at each depth for both cases are extracted and processed into flat clusters of loci, which are compared between the two cases using adjusted mutual information (AMI). **(B)** Results for every pair of domains identified at depth *i* and *j* with chromosome banding or HiDeF. The amount of AMI is multiplied by 100 and displayed on each cell. The *bottom-left* cell of each matrix corresponds to comparing the whole chromosome with itself (domain defined at the zeroth depth) which yields an AMI of 1 (*red cell*). The comparison between the whole chromosome and any other way of domain separation yields an AMI of 0 (*blue cells*).

Overall, there is an agreement between the chromosome bands and our predicted domains. The first two levels of the HiDeF hierarchy separate the mESC chromosome 5 into approximately

5-7 large clusters, which are broadly consistent with the five main chromosome bands A-G (band D only corresponds to a small region; see **Figure 2.13A**). Then we calculated the adjusted mutual information (AMI) as a metric to quantify the similarity between chromosome bands and HiDeF/GNM domains defined at different levels. The result corroborated the consistency between GNM/HiDeF domain separations at various levels and the experimental cytobands (**Figure 2.13**).

2.2.2.4 Enrichment of architectural binding proteins at the boundaries of deeply nested domains

To further understand the biological significance of the chromatin structural hierarchy, we compared our domain boundaries with the occupancy of several known architectural proteins in mammals (cohesion subunit RAD21; condensin subunits CAP-H2 and CAP-D3, CTCF, and transcription factor for polymerase III C (TFIIIC), see Section 2.2.4.1) in **Figure 2.14** for mESC chromosomes 5 and 10 and **Figure 2.14** for mESC chromosome 3.

As mentioned above, architectural proteins are enriched at the chromosomal domain boundaries and their density/occupancy correlates with the boundary strength (Gomez-Diaz and Corces, 2014; Van Bortle, et al., 2014). The highly packed tail mESC chromosome 5 (chr5:110,000,000-151,700,000, or the range 2050-3034 based on the axes' units of 50 kbs; **Figure 2.15B**), for example, appears to be enriched in architectural proteins as compared to the rest of the chromosome 5, consistent with the presence of a relatively high number of domain boundaries at that particular region. Precisely, the tail region exhibits a 15.4% increase in domain boundaries compared to the average density based on 536 boundaries over the entire sequence 151,700 kbs.

A close-up view of the region chr5:37,500,000-42,500,000 (or the range 750-850 in **Figure 2.15A**) further showed that the domain boundaries identified at different depths tend to co-localize with the architectural proteins. Based on the generated hierarchy, we could conveniently trace

specific domains to find out which *k* gives rise to their groupings. For example, in **Figure 2.15** panels **A** and **C**, the domains at depth 6 and 8 originate from GNM clusters obtained at modes $k \cong$ 100 and 300, which can be visually confirmed by the corresponding cross-correlation maps (**Figure 2.15C**).

Despite their distinct sequence lengths, different chromosomes exhibit hierarchies of similar depth (14.8±1.5, **Figure 2.16A**). Domain boundaries at all depths showed significant architectural protein enrichment as compared to the background (empirical *p*-value $\ll 0.0001$, **Figure 2.15D**), but more interestingly, the architectural proteins showed greater enrichment at higher (shallower) levels of the hierarchy than at lower (deeper) levels, suggesting that the deeper domain boundaries tend to be weaker than the shallower ones. This may be because the domain identified at a deeper level are packed and wrapped around by their upper level structures, such that they can be either stabilized by fewer architectural proteins or they are less exposed to them, or both. Alternatively, a more flexible definition of domain boundaries on a local scale may be necessary for facilitating the gene transcription regulation. Since deeper domains tend to have smaller sizes while the sizes of domains at the same depth may vary, we checked whether the correlation between the architectural protein binding and the domain depth is simply an effect that can be attributed to the domain sizes. But surprisingly architecture protein binding occupancy did not show any correlation with domain sizes (r = -0.04, **Figure 2.16B**).

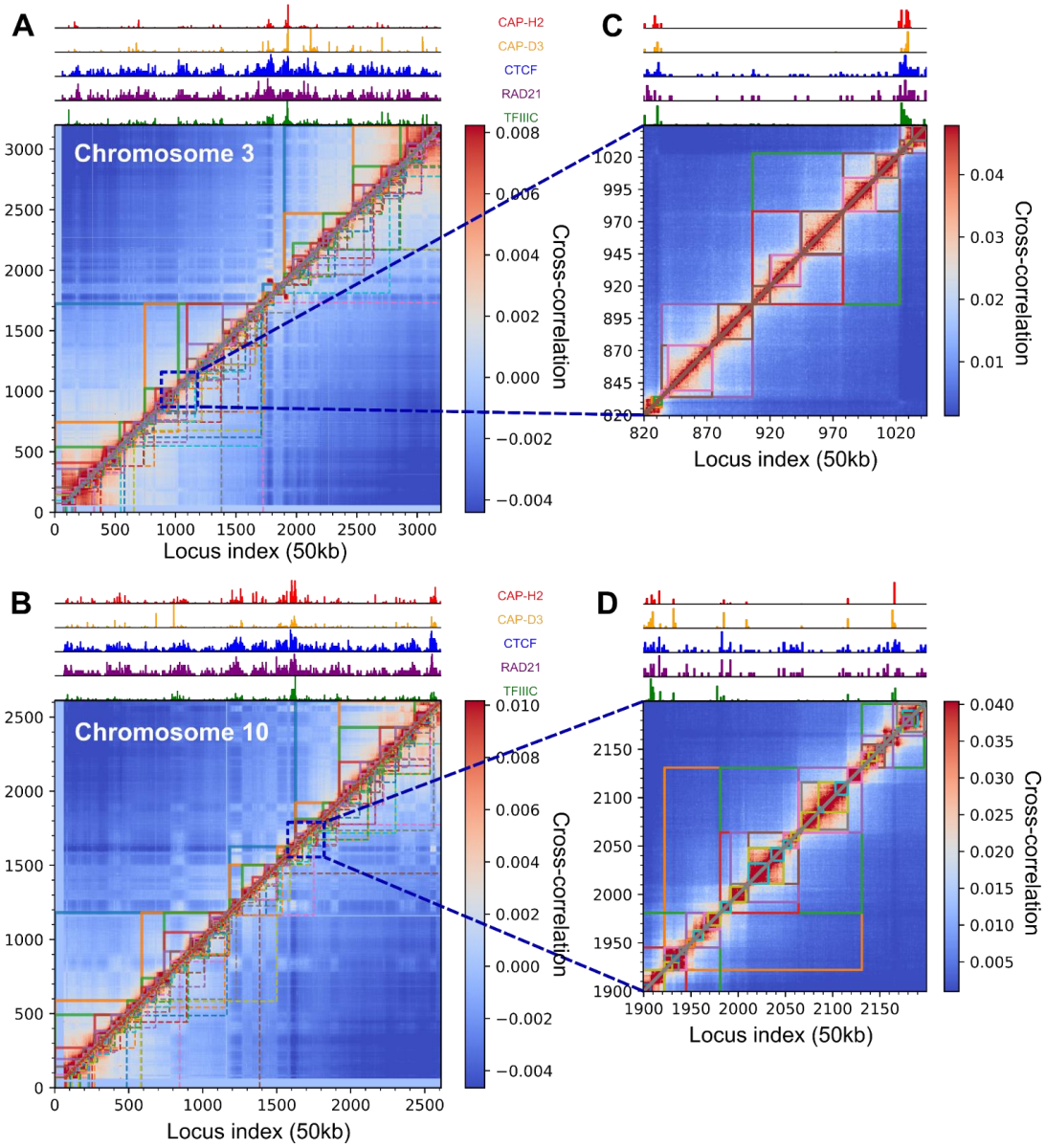

Figure 2.14 Hierarchical spatial organization of mESC chromosomes 3 and 10 and comparison with the loci of architectural proteins. (A) Cross-correlations calculated for chromosome 3 using all GNM modes. Similar representations of cross-correlations and domains are adopted as in **Figure 2.12**. *Colored lines* in the *upper triangle* delineate the domain boundaries redefined based on the hierarchy and those in the *lower triangle* delineate the original domain boundaries identified the GNM modes. The histograms on *top* of the matrix show the architecture protein occupancies. RAD21: cohesion subunit; CAP-H2 and CAP-D3: condensin subunits; CTCF: CCCTC-binding factor; TFIIIC: transcription factor for polymerase III C. **(B)** Same results for chromosome 10. **(C)** A close-up view of the results at genomic regio chr3:41,000,000-52,250,000. **(D)** Close-up at genomic region chr10:95,000,000-110,000,000.

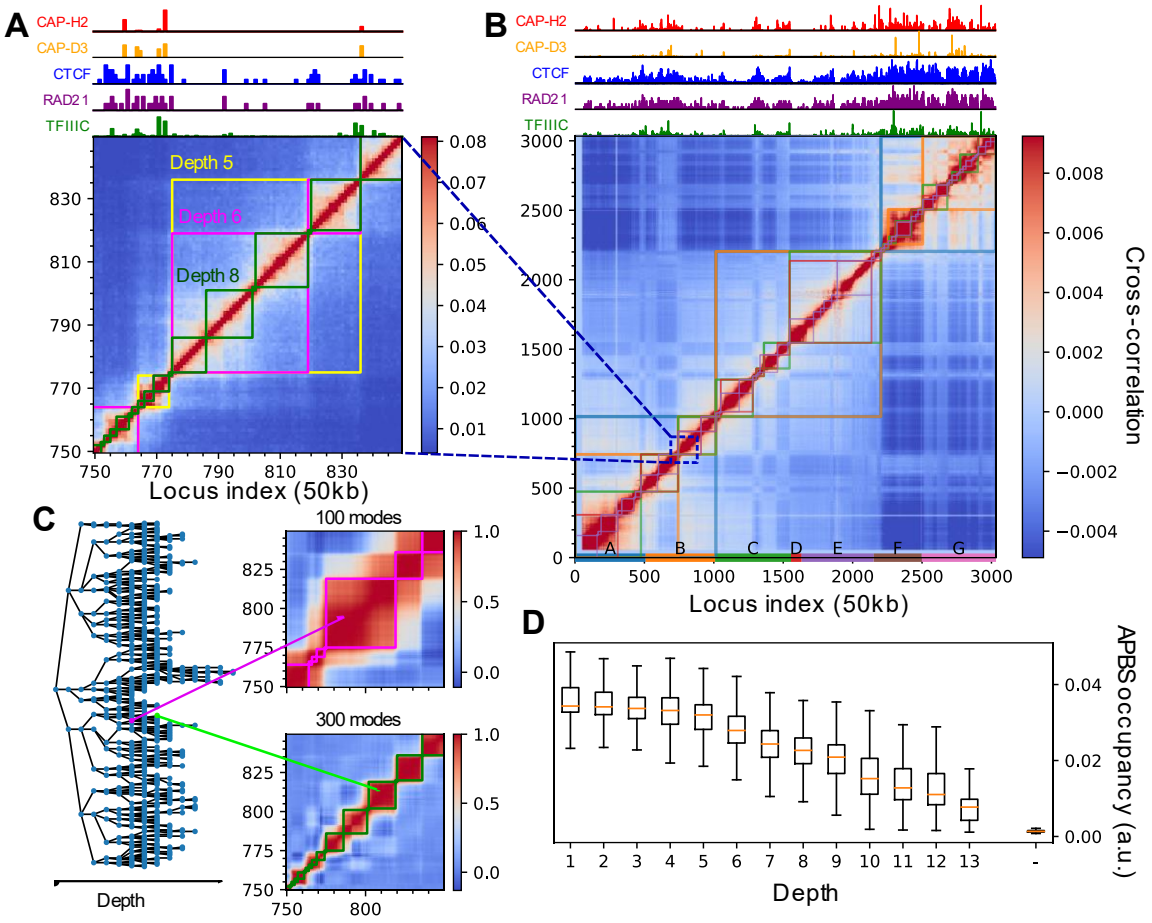

Figure 2.15 Hierarchical organization of mouse chromosome 5. (A) A close-up view of cross-correlations calculated for mESC chromosome 5 using all GNM modes analyzed by HiDeF (chr5: 37,500,000-42,500,000; or loci 750-840, each locus being composed of 50kb). *Colored lines* delineate the boundaries and sizes of domains identified at different depths. The histograms on *top* of the matrix show the occupancies of architectural proteins at corresponding loci. **(B)** Same results for the entire chromosome 5. **(C)** Hierarchy of chromosomal domains identified by HiDeF/GNM. The matrices on the right are cross-correlations calculated from 100 (*top*) and 300 (*bottom*) GNM modes. Note the finer-grained distribution of correlated regions as we proceed to higher modes. The *magenta* and *green boxes* correspond to the two depths indicated by the *colored arrows* in the matrices. **(D)** Normalized distributions of architectural protein binding site (APBS) occupancies within a two-locus radius around domain boundaries identified at different depths. The last entry ("-") along the abscissa represents the background APBS occupancy. The boundaries from earlier depths are excluded in latter depths. The zeroth depth (the entire chromosome) and depths that have <500 boundaries are omitted.

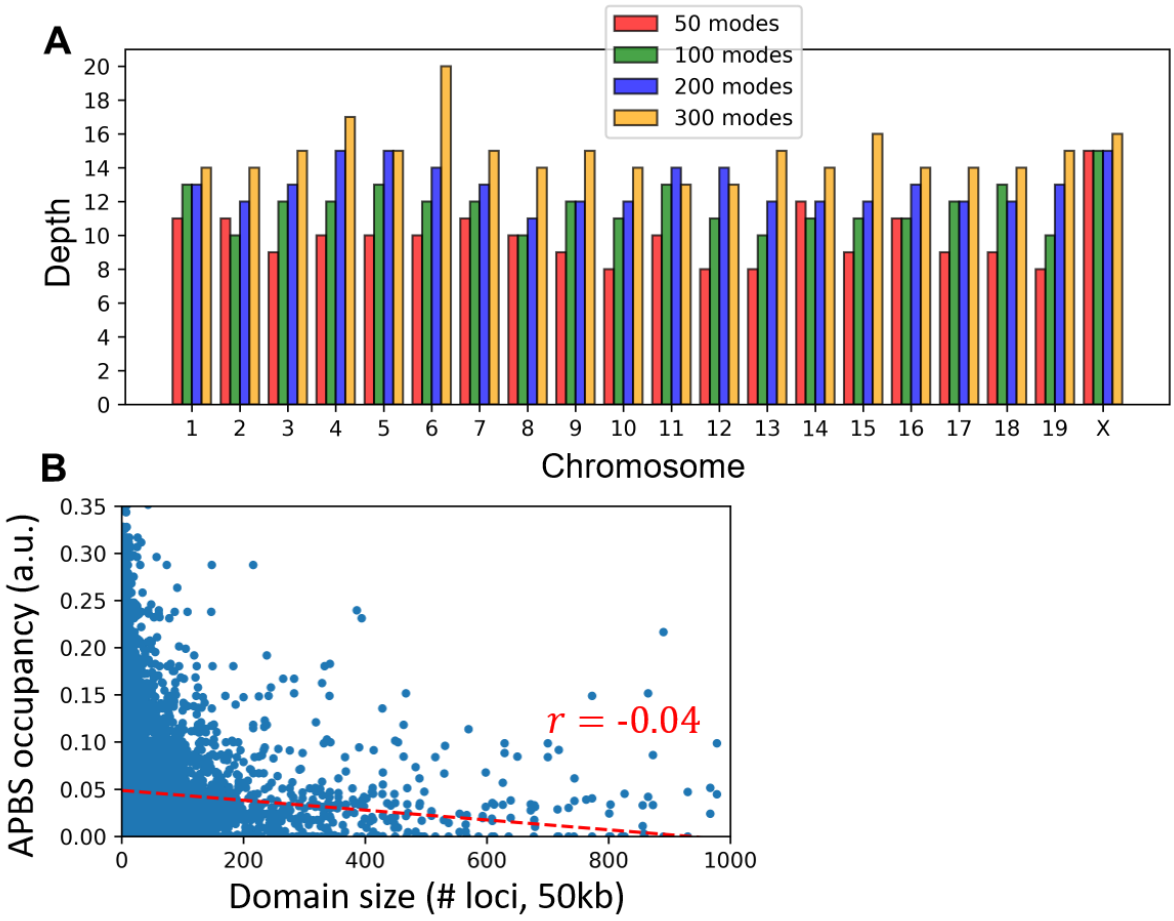

Figure 2.16 Invariants of the hierarchical organization of chromosomes. (A) Depths of hierarchies extracted for different chromosomes using clusters identified at different *k* (number of GNM modes). The depth *i* ncreases with *k*, but not as fast. In addition, the depth of the hierarchies obtained for a given *k* does not depend on the size of the chromosomes (even though larger chromosomes have intrinsically access to a larger number of modes). **(B)** Relationship between APBS occupancy and the domain size. Each *dot* represents an architectural protein binding occupancy (*ordinate*) at the boundary of a domain with some size (*abscissa*). The domain size is quantified as the number of loci, and since each boundary borders two domains, the domain with bigger size is selected. The *red line* is the best linear fit (r = -0.04, with $p < 0.001$).

2.2.3 Discussion

In the Subsection 2.2, we extracted and analyzed the hierarchical organization of chromosomal regions, the size of which varies from whole chromosome arms to small genomic loci, including compartments, subcompartments, TADs, and other chromosomal domains at intermediate levels/depths. We used a novel hierarchy inference framework, HiDeF, in combination with the GNM identification of structural domains. We showed that the domains extracted from the GNM modes correlate with known chromosomal structures such as compartments and TADs; the genomic regions in different subtypes of compartments exhibit differentiated chromatin mobility and accessibility; the depth of a chromosomal region in the hierarchy may be an important structural feature that can be revealed by its boundary strength. Alternatively, depths with strong boundaries inform which domains exhibit robust structural and dynamic coherence. In the future, a similar analysis could be used for comparing the chromatin structure across different cell types or even species.

The GNM identification of structural domains is mathematically related to spectral clustering through the common Laplacian/Kirchhoff matrix (see Appendix A). We note that a TAD-finding method using only the second eigenpair (Fiedler value/vector) of the Laplacian has also been developed (Chen, et al., 2016) and tested on 100 kb resolution data. By including a higher number of eigenvectors, we were able to identify TADs comparable to those detected by Armatus on all chromosomes (as measured by lower VI) at 5 kb, and for 18/23 chromosomes at 100 kb resolution (data not shown; see Supplementary Figure S11A and C in (Sauerwald, et al., 2017)). Further corroborating the benefit of using multiple modes, earlier studies showed that spectral clustering by using more eigenvectors can outperform partitioning methods which only use one eigenvector (Alpert, et al., 1999; Alpert and Yao, 1995). In addition, we note that a TAD calling

program, TADtree, also returns hierarchical domains, but it exclusively identifies domains at the scale of TADs (~1Mb) (Forcato, et al., 2017; Weinreb and Raphael, 2016); whereas in the currently proposed approach, mega-domains of up to 100Mb can be identified, alongside TADs or regions of different sizes. As a result, the hierarchical domains reported here tend to be more nested than those identified by TADtree.

Multiresolution community and hierarchy detection methods have been of broad interest in recent decades (Blondel, et al., 2008; Reichardt and Bornholdt, 2006; Rosvall, et al., 2009). While most methods for constructing the hierarchies take recursive bottom-up or top-down approaches, here, we proposed an alternative approach of a modular workflow, which decouples the inference of clusters from the identification of their hierarchical relations. Such modularity makes it possible to substitute alternative algorithms at each step, so one could use clustering algorithms other than the GNM/spectral clustering as the basic method for identifying domains at different resolutions, and then combine the results with HiDeF algorithm to robustly extract the underlying hierarchy. Therefore, the proposed framework would be expected to be of broad utility for analyzing multiscale network organization in many research domains. For instance, it could be used for revealing hierarchical cell-type relationships in single-cell RNA-seq data to gain insights into the composition and development of tissues/organs (Stuart, et al., 2019); or for discovering protein complexes or modules from comprehensive protein-protein interaction (PPI) databases (Li, et al., 2017) or high-throughput screens of PPIs (Szklarczyk, et al., 2019) for aggregating genomic information of diseases and inferring novel disease genes.

2.2.4 Methods

2.2.4.1 Hi-C and ChIP-seq data processing

Data were obtained from the Gene Expression Omnibus (GEO) database (Barrett, et al., 2013). Population Hi-C data for human GM12878 cell line came from GSE63525 (Rao, et al., 2014) and those for mESCs were obtained from GSE80280 (Stevens, et al., 2017). Both datasets are normalized using VCnorm (Rao, et al., 2014). ChIP-seq data for architectural proteins were obtained from GEO series GSE90994 (CTCF), GSE33346 (cohesin: RAD21; condensin subunits: CAP-H2 and CAP-D3), and GSE80034 (TFIIIC) (Cattoglio, et al., 2019; Dowen, et al., 2013; Yuen, et al., 2017). Processed data files were downloaded and converted to BED format using BEDOPS (Neph, et al., 2012). Chromosome banding data were obtained from Ensembl genome browser (Yates, et al., 2020). Cytobands were parsed as label arrays and the underlying hierarchies were built using HiDeF. Genome positions in different assemblies were converted to those in GRCm38/mm10 with liftOver in the UCSC genome browser (Kent, et al., 2002). All data were normalized and binned into 50kb per loci. Architecture protein binding site (APBS) data were normalized for each chromosome to obtain occupancies. Those located within a two-locus radius around domain boundaries were collected and categorized based on the depth of the domains, and those outside, were considered as the background. To balance the sizes, data belong to each depth and the background were replicated with 1,000 bootstrap samples of size 100. We used Welch's t-test to quantify the difference between the distribution of APBS occupancies of each depth and the background.

2.2.4.2 Variation of Information (VI) metric

This metric is based on information theory. It measures the difference in information contained in two clusterings, or partitions, of a dataset. If we consider each domain to be a cluster of nodes/points, this type of comparison becomes very natural. Formally, for two sets of clusters *A* and *B*, VI is defined as follows:

$$VI(A,B) = H(A) + H(B) - 2I(A,B) \tag{2.7}$$

where $H(A)$ represents the entropy of a set of clusters A, and $I(A,B)$ is the mutual information between the two partitions, given by

$$H(A) = -\sum_{i} P(i)\log P(i)\,, \tag{2.8}$$

$$I(A,B) = \sum_{i,j} P(i,j)\log\frac{P(i,j)}{P(i)P(j)}, \tag{2.9}$$

where the probability of picking a node in cluster C_i, $P(i)$, is simply the number of points in that cluster divided by the total number of points in the data set. In this work, a "cluster" is the set of loci placed into the same domain or compartment.

Note that this is a true metric in the space of clusterings; VI is commutative, satisfies the triangle inequality, and is always non-negative and equal to zero if and only if the two clusterings are identical. More intuitively, VI is a measure of the amount of information that is lost and gained by changing from one clustering to another, without any assumptions placed on the clusterings themselves or how they were generated. More information can be found in (Meilă, 2003).

2.2.4.3 Multi-resolution spectral clustering

We used spectral clustering techniques to identify chromosomal domains from the GNM modes (Appendix A). The GNM-predicted intrinsic dynamics is defined by a spectrum of normal modes of motion. LF modes usually correspond to global motions, i.e. they collectively engaging large domains; whereas HF motions refer to local movements (of small domains or individual loci). Therefore, the frequencies of the GNM modes can be regarded as a "resolution parameter" that controls the size of domains engaged in collective motions. For each chromosome, we calculated the GNM modes as well as the cross-correlations between the motions of gene loci, based on the normalized Hi-C contacts as previously described (see Section 2.1.4.2) and identified genomic loci clusters by discretizing the first $k = 1, 2, \ldots, 300$ modes (Stella and Shi, 2003). This step led to 300 sets of clusters per chromosome. Loci in the same cluster that were not sequentially consecutive were separated into different domains (**Figure 2.17**). We found that the results from discretization were robust to randomization (Stella and Shi, 2003); therefore, we presented in this study the results from one single run. Multiple runs were performed and yielded similar results. Variation of information (VI) metric was used as a quantitative measure of agreement between GNM-predicted domains, TADs, and compartments.

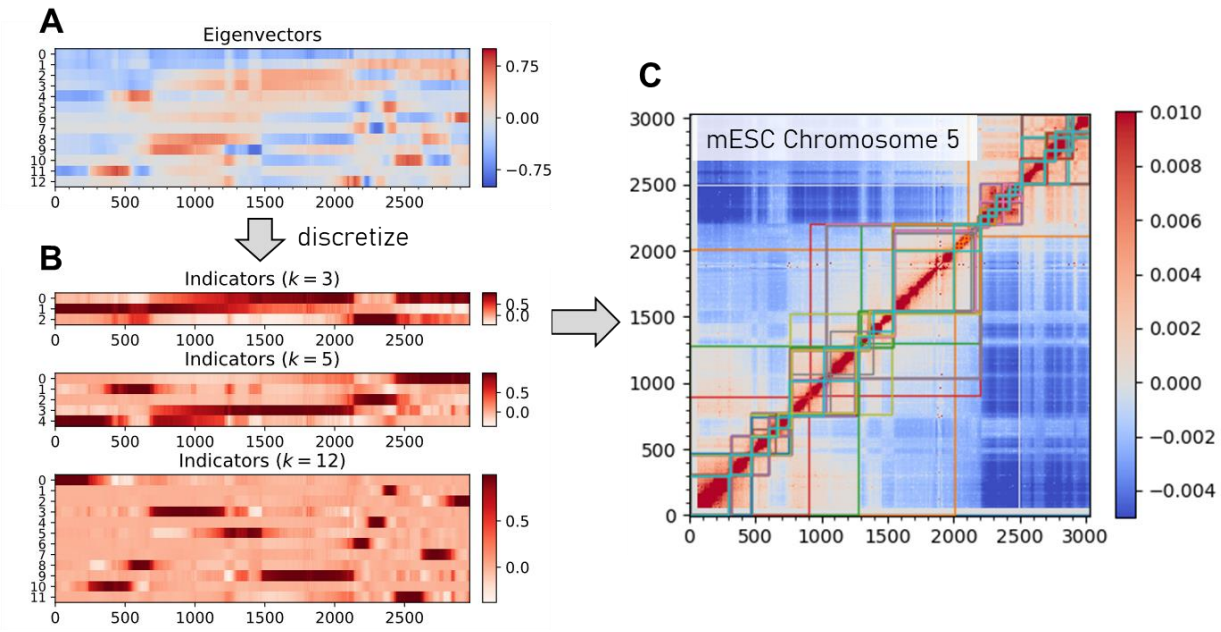

Figure 2.17 Schematic shows how GNM modes (eigenvectors) are discretized and processed into indicators for finally identifying the GNM domains. (A) Eigenvectors solved from the GNM and sorted ascendingly based on associated eigenvalue. Each row represents an eigenvector whose positive/zero/negative elements are indicated by *red*/*white*/*blue* colors. **(B)** Indicator vectors discretized from first *k* eigenvectors whose elements are nonnegative. **(C)** The resulting multi-resolution chromosomal domains. Domains separated based on different *k* are outlined with different colors.

2.2.4.4 Inferring hierarchical organization of chromatin structure

The containment relationship between two chromosomal domains, A and B, is quantified by the containment index as:

$$ci(A,B) = \frac{|A \cap B|}{|B|}, \tag{2.10}$$

which intuitively measures how much of B is shared with A. A *containment graph* can be constructed based on the measure, where vertices represent clusters identified from the previous step (see Section above) and edges represent containment relationships. Specifically, given a cutoff value $\kappa \in (0.5, 1]$, if $ci(A,B) > \kappa$ and $ci(A,B) > ci(B,A)$, then there exists an edge from A to B,

representing that *A* κ-contains *B*. A *root* vertex is added to connect to all other vertices, representing that every data point belongs to a grand hypothetical cluster. A κ value of 0.9 was used to generate results in the present study.

The containment graph is a DAG that denotes a complete, albeit redundant, set of containment relations, meaning that if *A* κ-contains *B* and *B* κ-contains *C*, it is likely (but not necessary) that *A* also κ-contains *C*. Redundant relations/edges are removed by obtaining a *transitive reduction* of the original graph. In addition, edges are considered as competing if they are incident to the same vertex (i.e. biological pleiotropy), and a second cutoff value, ξ, is used to control how many competing edges are tolerated in the final graph. Vertices are labeled with numerical depths equal to their maximum distances to the root. Clusters/domains are redefined at different depths conforming to the structure of the graph/hierarchy (see details in Appendix B). ξ = 0.2 (tolerating top 20% competing edges) was adopted in the present study (competing edges were omitted in plots for clearer visualization).

2.2.5 Acknowledgment

Part of this work was published previously (Sauerwald, et al., 2017). Natalie Sauerwald contributed to the identification of GNM domains and performed comparative analysis with TADs and compartments. Dr. Ivet Bahar and Dr. Carl Kingsford supervised the project.

The development of HiDeF is a collaborative work between Drs. Trey Ideker, Fan Zheng, Ivet Bahar and me. I led the design and development of HiDeF and F.Z. made significant contributions. I.B. and T.I. supervised the project.

Special thanks to Dr. Chakra Chennubhotla, Dr. Luca Ponzoni, Dr. Burak Kaynak, and Yan Zhang for useful discussions and advices on the topic of spectral clustering.

3.0 Comparative Study of Chromosomal Dynamics of Different Cell Types toward Understanding Cell-to-Cell Heterogeneity

Cell identity is determined by lineage-specific gene expression during differentiation (Bernstein, et al., 2007). The process of gene expression is regulated by the accessibility of the corresponding region of the DNA to transcription factors and co-factors. Therefore, the spatial organization of the genome plays a crucial role in the process of cell differentiation. Recently, various studies have shown that different cell types show recognizably different contact topologies at their chromatin and the observed changes in the chromatin structure are associated with cell development and differentiation (Andrey and Mundlos, 2017; Bonev, et al., 2017; Dixon, et al., 2015; Joeng, et al., 2017) (**Figure 3.1**). However, questions remain regarding the type and extent of conservation and/or differentiation of chromatin structure among different cell lineages and how to quantify these differences.

Advances in chromosome conformation capture techniques as well as computational characterization of genomic structural dynamics open new opportunities for exploring the structural aspects of genome-scale differences across different cell types. Rao et al. (Rao, et al., 2014) found that many loop domains (~100 kb) are conserved not only in different cells but also across species; Dixon et al. (Dixon, et al., 2015) noted that, although chromatin domain boundaries tend to be stable during cell differentiation, drastic changes in chromatin interactions are observed both within and between domains; Rudan et al. (Rudan, et al., 2015) found that the CTCF sites, one of the most important determinants of domain boundaries, evolve under two regimes: some CTCF sites are conserved across species, others are significantly more flexible. A recent single

cell study on mammalian genome showed that while larger chromatin structures compartments are mostly conserved, the structures of the topologically associating domains (TADs) and loops may vary substantially even within the population of the same type of cells (Stevens, et al., 2017). All these observations have shown some levels of conservation as well as variation in the chromatin 3D structure or organization of different cells, suggesting a complex dependency on cell type at the 3D genome level, which is further obscured by cell heterogeneities within even a given type of cell.

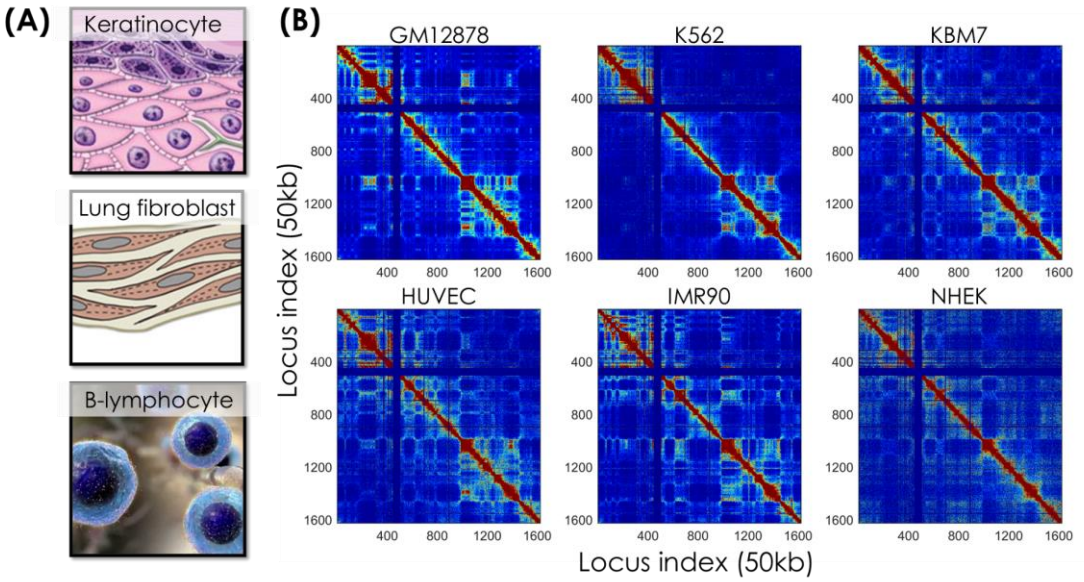

Figure 3.1 Heterogeneity in chromosomal spatial organization across different types of cells. (A) Diagrams of three cell types derived from different germ layers. **(B)** Contact maps of chromosome 17 measured by Hi-C experiments for six different cell types (Rao, et al., 2014): GM12878, human B-lymphocytes; K562, human immortalized myelogenous leukemia line; IMR90, human lung fibroblasts; NHEK, primary normal human epidermal keratinocytes; HUVEC, human umbilical vein endothelial cell line; KBM7, chronic myelogenous leukemia cell line.

The models, methods and tools presented in the previous sections showed that chromosomal structural dynamics is an important feature that explains/defines cell identity, in

addition to other epigenetic properties. Comparison with RNA-seq expression data will reveal in the present chapter (also published in (Zhang, et al., 2020)) a strong overlap between highly expressed genes and those distinguished by high mobilities in a cell-specific manner, in support of the role of the intrinsic spatial dynamics of the chromatin as a determinant of cell differentiation. Overall, network models for characterizing 4D genome dynamics provide a computationally efficient platform for assessing cell type-specific behavior and differentiation at the level of the entire chromatin.

3.1 Conservation *vs* Variation of Chromosomal Dynamics

3.1.1 Introduction

We examined here the dynamic basis of variabilities between different cell types by investigating their chromatin mobility profiles inferred from Hi-C data using an ENM representation of the chromatin. As explained and shown in Section 2.1.2.1, the MSFs of network nodes (i.e. loci in the context of chromatin) positively correlate with their accessibility: mobile sites are more likely to be exposed than stationary ones, and therefore genes that located at such sites tend to be more accessible to transcription factors and other regulatory proteins. Thus, it may be of interest to compare MSFs of chromosome loci across different cell types to identify cell type-specific exposed chromosomal regions toward understanding the structural and dynamic bases of differential gene expression.

MSFs are obtained from the linear (weighted) combination of square displacements of loci, contributed each by the full set or a representative subset of normal modes. As explained in Section

2.1.4.3, the *shape* of the mode (eigenvector) describes the direction and magnitude of the intrinsic collective motions of the loci, and the *frequency* (eigenvalue) serves as a weight to the mode that defines the amplitude of the (square) displacement along the mode. Modes are assigned increasing *mode numbers* with decreasing amplitude, such that the first few modes (global modes) make relatively large contributions to the fluctuation spectrum; whereas higher modes usually exhibit more localized fluctuations. As a consequence, the differences we see in MSFs are twofold: those originating from the differences in the mode shapes and those resulting from the differences in the frequency dispersion of the modes. Therefore, the evaluation of mode spectra is not only a necessary step for computing overall dynamical quantities such as the MSFs, but also a useful tool for assessing the origin of the differences in the spatial dynamics of the genomes of different cells.

In this section, our comparative analysis of sixteen cell lines reveals close similarities between chromosomal dynamics across different cell lines on a global scale, but notable cell-specific variations emerge in the detailed spatial mobilities of genomic loci. Closer examination of the mode spectra reveals that the differences in spatial dynamics mainly originate from the difference in the frequencies of their intrinsically accessible modes of motion. Thus, even though the chromosomes of different types of cells may have access to similar modes of collective movements, not all modes are deployed by all cells, such that the effective mobilities and cross-correlations of genomic loci are cell type-specific.

3.1.2 Results

We first evaluated the GNM modes and predicted MSFs of genomic loci for all chromosomes in16 human cell lines (**Table 3.1**), using their inter-loci contact topology data from public Hi-C datasets (Darrow, et al., 2016; Joeng, et al., 2017; Phanstiel, et al., 2017; Rao, et al.,

2014; Rao, et al., 2017; Sanborn, et al., 2015) in the GNM framework extended for chromatin dynamics as described in Section 2.1.4.3. MSFs were analyzed and compared across different cell lines to assess the overall similarities between the mobility profiles of the individual chromosomes. Next, similar to the comparative analysis we did for protein families in Section 1.2, we matched and examined equivalent normal modes across cell lines and assessed the extent of mode-mode overlaps and the origin of the variations in loci mobilities.

3.1.2.1 Genomic loci exhibit similar fluctuations on a global scale while retaining cell-specific patterns

We first examined the MSFs of genomic loci evaluated for different cell types. Mobility profiles (normalized MSFs, which allow for visual comparison of the behaviors of different cells) are illustrated for the chromosomes 17 (**Figure 3.2A**) and 2 and 8 (**Figure 3.3**) of the 16 different types of cells. The series of curves in these figures show chromosome-specific patterns broadly shared by different types of cells. Cell-cell similarities between chromosomal mobility profiles are quantified by pairwise Pearson correlation coefficients. This led to an average Pearson correlation of $r = 0.63 \pm 0.23$ for cell-cell mobility profiles of all chromosomes (*dashed line* in **Figure 3.2B**), except for an ectodermal cell line, NHEK, which correlated poorly with other cells ($r = 0.29 \pm 0.22$, **Figure 3.2B**) but exhibited a correlation with another ectodermal cell line, RPE1 ($r = 0.58 \pm 0.14$). RPE1, in turn, exhibited relatively strong correlations with two other ectodermal cell lines, HMEC ($r = 0.66 \pm 0.08$) and HCT116 ($r = 0.63 \pm 0.12$), as well as the only endodermal cell line, IMR90 ($r = 0.68 \pm 0.11$).

Pearson correlations between the mobility profiles of genomic loci in different types of cells can be viewed in **Figure 3.2C** for all pairs of cells. The results refer to the collective fluctuations of all chromosomes for each pair of cell types. The hematopoietic cell lines (the first

nine in **Figure 3.2B-C**) exhibit a correlation of at least 0.5 with each other, with three being the most dissimilar, EP (erythroid progenitors), THP1 and THP1-derived macrophages. EP is the only red blood cell line in the dataset, and THP1 and the macrophages are more differentiated than other hematopoietic cell lines. HSPC, the hematopoietic stem and progenitor cells, correlate well with almost all other cell types (r = 0.66 ± 0.17, **Figure 3.2B**, first entry) and will be used as reference for quantitative analyses of cell type-specific behavior.

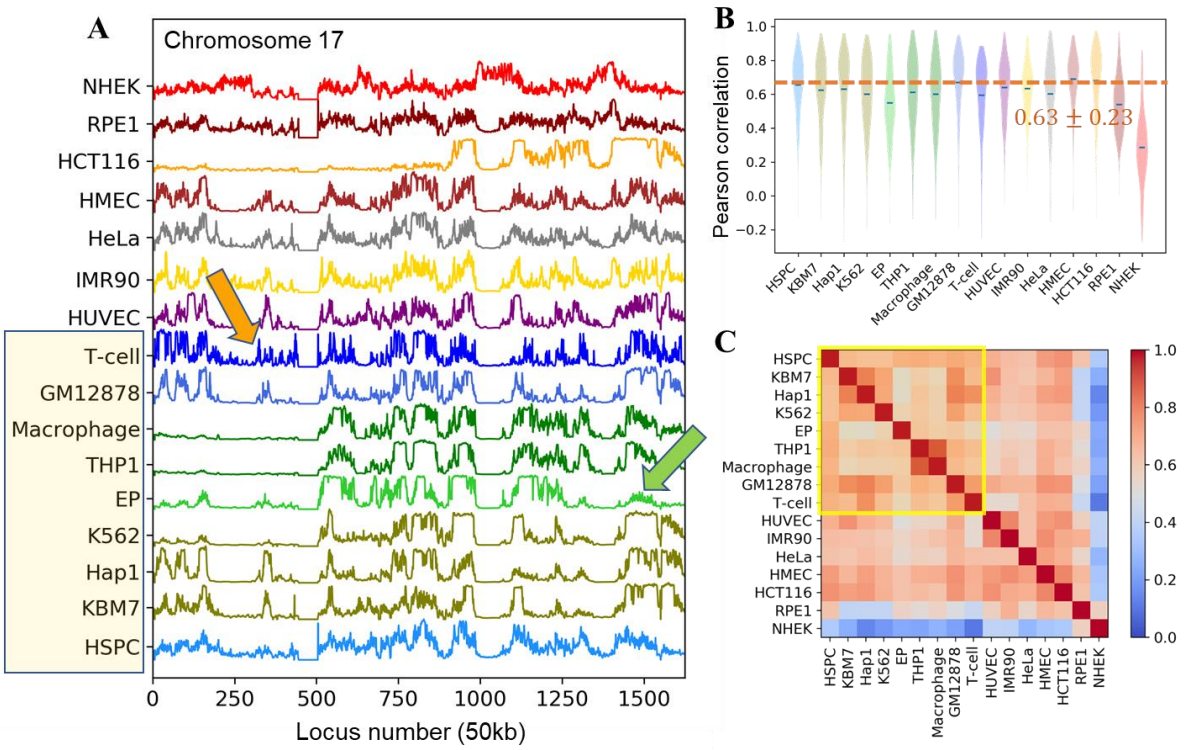

Figure 3.2 Comparison of the chromosomal dynamics of different types of cells. (A) Mobility profile of genomic loci computed for the chromosome 17 of all 16 cells or cell lines in our dataset (**Table 3.1**). The *curves* represent the normalized distributions of MSFs of gene loci predicted by the GNM, stacked up for visual comparison. **(B)** *Violin plots* showing the distribution of Pearson correlations between the chromatin mobility profile of individual cells (listed along the *abscissa*) and all others. Results are computed using the first 500 matched modes predicted for the entire chromatin of all cells. *Blue dashes* indicate the mean. **(C)** Heat map showing the Pearson correlations between the intrinsic dynamics of the examined 16 cell lines, based on the mobility profile of all chromosomes. This figure is adapted from (Zhang, et al., 2020).

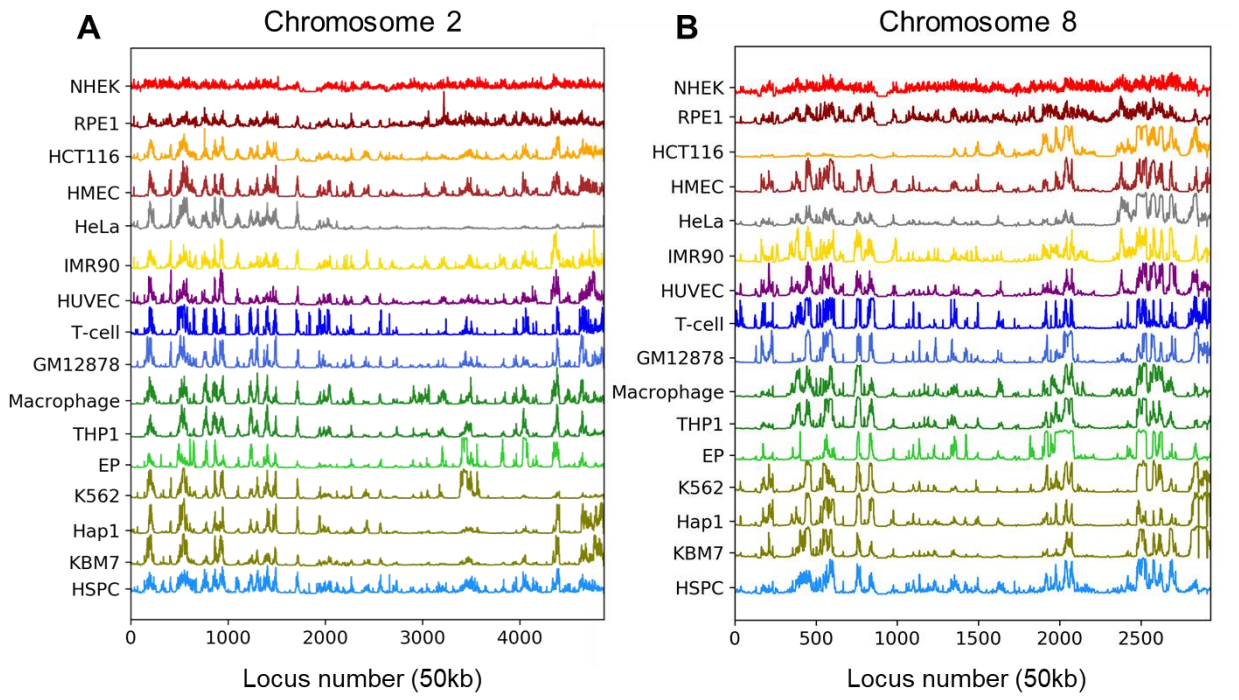

Figure 3.3 Mobility profiles for two chromosomes computed for 16 types of cell lines. Results are shown for chromosomes 2 **(A)** and 8 **(B)**. Mobility profiles of genomic loci were computed based on first 500 GNM modes. This figure is adapted from (Zhang, et al., 2020).

3.1.2.2 Dissection of mode spectra reveals the high conservation of global modes

Decomposition of the chromosomal mode spectrum accessible to each cell type yielded the *mode conservation* curve presented in **Figure 3.4A** as a function of mode number. The ordinate $\langle S \rangle_i$ designates the correlation cosine between the shape of mode *i*, averaged over all pairs of cells (see Section 3.1.4.2). The first mode is highly conserved across all examined cells ($\langle S \rangle_1 = 0.84 \pm 0.18$) indicating the prevalence of a *global mode shape* for the chromatin, shared by all cell lines; **Figure 3.4C** and **D** illustrate the global mode shape (parts) corresponding to the respective chromosomes 2 and 17. A closer look at mode conservation within individual chromosomes (illustrated for chromosome 2 in **Figure 3.5E**, for example) also exhibited the same pattern, mainly

high-to-moderate conservation of the first few modes (see also the *inset*, **Figure 3.4A**), followed by a rapid decrease with increasing mode number.

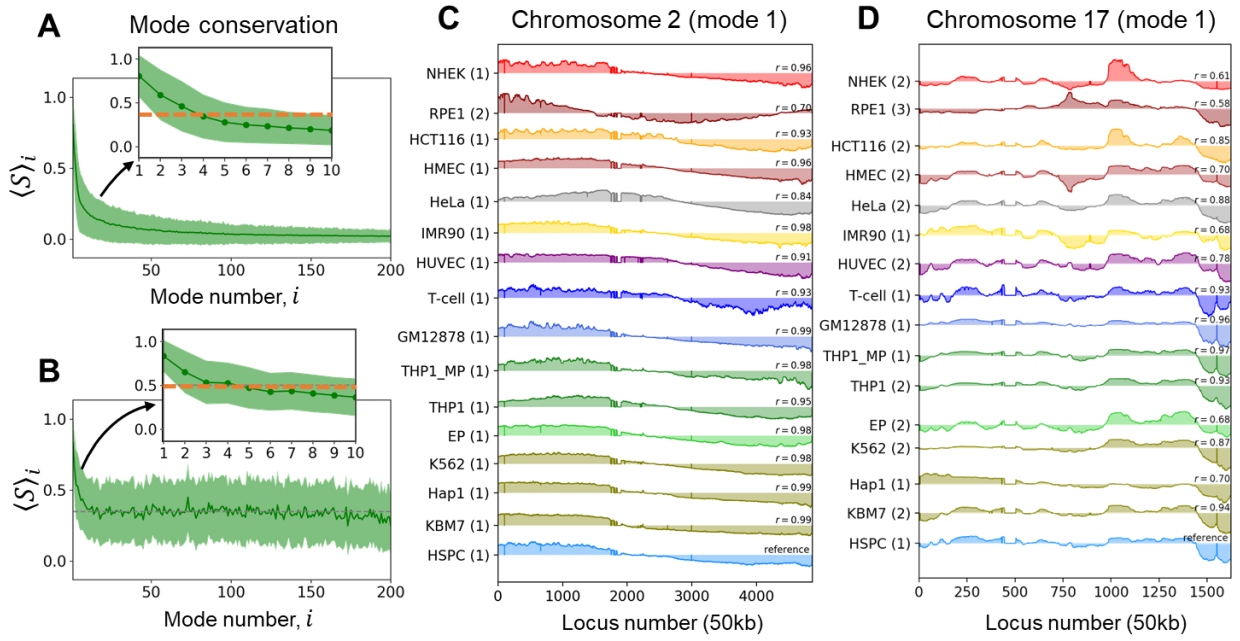

Figure 3.4 Conservation of the first global mode accessible to the chromatin. (A) Mode conservation profile as a function of mode number i reveals the high conservation of the first global mode (i = 1). The profile displays the correlation cosines between individual mode shapes computed for each of the first 200 modes accessible to the entire chromatin averaged over all cell pairs (equation 3.2). The *solid green curve* and the *shade* show the mean and standard deviation, respectively. **(B)** Same as panel **A**, after reorganizing the modes to select the equivalent modes that best match those of the reference cell, HSPC. Note the higher conservation of modes, but also the accompanying higher variance. **(C-D)** Global mode shape for chromosomes 2 (**C**) and 17 (**D**), highly conserved across 16 cell lines. The *curves* represent the normalized spatial displacements of loci (*abscissa*) along the equivalent mode 1 axis. A central hinge region is observed in **C** at the crossover between positive and negative displacements near loci 2,000-2,500. The original mode numbers are shown in parentheses on the *ordinate*, and the correlation cosine with respect to the reference (HSPC) is indicated in each case. This figure is adapted from (Zhang, et al., 2020).

We further evaluated the mode-mode overlaps among the first 10 modes, for every pair of cells. The heatmap on the *left* in **Figure 3.5A** shows an example of such overlaps for HSPC and

KBM7, where each element represents the correlation cosine $[S(A,B)]_{ij}$ (equation 3.1) between the i^{th} and j^{th} modes ($1 \leq i,j \leq 10$) of the respective cells A and B (in this case HSPC and KBM7). The same type of overlap map is displayed for HSPC and each of the other cell types in **Figure 3.5C**. These maps confirmed that the slowest modes from different cells exhibit relatively high overlaps (see *red pixels* near the *upper left* part of the diagonal in each block). Even in the case of the most distinctive cell types such as GM12878 and NHEK, the overlap between the top three modes remained above 0.65.

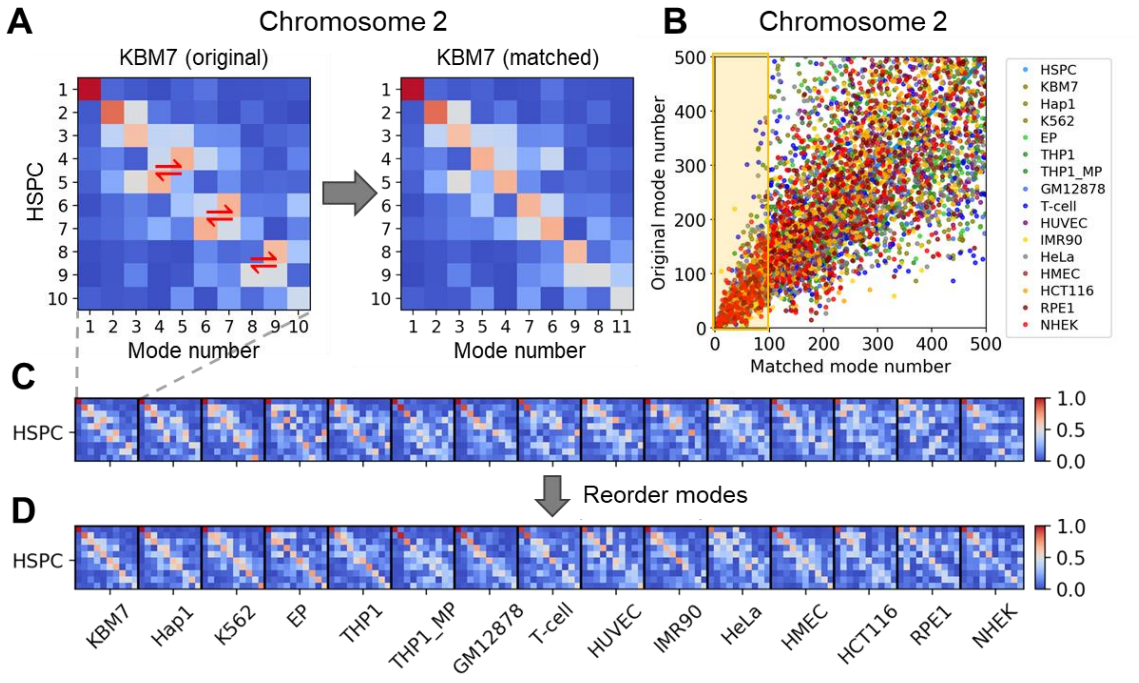

Figure 3.5 Mode-mode overlaps across different cell lines, illustrated for chromosome 2. (A) Mode-mode matching process illustrated for the softest 10 modes of HSPC and KBM7. The entries in the heat map are the mode-mode correlation cosines, with the strength of correlations decreasing from *red* to *blue*. *Red two-way arrows* display the modes that are swapped to result in the map on the right. **(B)** Comparison of the original (*abscissa*) and reassigned (*ordinate*) mode numbers after optimal matching to the mode numbers of the reference cell (HPSC). Results for different cell types are color-coded, consistent with previous figures. **(C-D)** Same as panel A, displayed mode-mode overlaps between all 15 pairs of cell types and the reference HPSC, and their reordering to identify equivalent modes. This figure is adapted from (Zhang, et al., 2020).

3.1.2.3 While different cell types have access to conserved genome-scale dynamics, the active modes of motions differ from cell to cell

Closer examination of the heat maps such as **Figure 3.5A** (*left*) reveal that a high mode-mode overlap between different cell lines is not necessarily observed at the diagonal elements of each block, indicating a mismatch in mode numbers between different cells. As mentioned earlier, the *mode number* is a physically meaningful quantity, smaller indices referring to lower frequency or larger amplitude modes. Thus, an off-diagonal *red pixel* in the heat map means that the two modes are similar in shape (relative distribution of loci movements during this mode), but not in size (absolute amplitude of motions). In a sense, the mode will be more pronounced or active in one type of cell compared to the other. Here, "more active" means a predisposition to undergo a relatively larger displacement along that mode (exhibited by the cell with the smaller mode number).

The differences in the mobility profiles of chromosomal loci in different cell types (**Figure 3.2A**) can thus originate not only from the different shapes of the modes - evidenced in the comparison of the global mode shapes of chromosomes 2 and 17 for the 16 cells/cell lines in **Figure 3.3C** and **D**), but also from their different frequencies or statistical weights.

To understand to what extent the frequency dispersion or the selective activation of pre-existing shared modes underlies the differences in the observed spatial mobilities of genomic loci, we adopted the mode numbers of HSPCs as reference (as the most undifferentiated cell in the dataset, based on the mode shape overlaps) and reordered the modes of the other 15 cell lines to achieve the highest mode-mode overlaps. **Figure 3.5A** provides a schematic description of mode number reassignment method. Essentially, the shape and frequencies of the modes are retained, but their mode index is changed to match the so-called "equivalent" modes in evaluating the

average mode-mode overlaps. This led to heat maps with highest mode-mode overlaps along the diagonals of the blocks, as illustrated in **Figure 3.5D**, and a conservation profile presented in **Figure 3.4B** for the entire chromatin. By selectively including the "equivalent modes" (or *matched* modes) and excluding others to evaluate the intrinsic dynamics, we end up with mobility profiles that are almost identical across all cell lines (**Figure 3.6A**). The recomputed mobility profiles led to significant increase in the Pearson correlations among different cell lines (r = 0.85 ± 0.08, **Figure 3.6B**; compared 0.63 ± 0.23 in **Figure 3.2A-B**).

It is important to note that the equivalent modes were identified by searching a broader range of modes, and often found from amongst "higher" modes (**Figure 3.5B**), which means some of the matched modes had relatively low weights/amplitudes and thus might not be contributing to collective dynamics in a given cell type as effectively as they do in another cell type. Slow modes tended to be retained without significant change in mode numbers; whereas fast modes exhibited large differences. For example, the first 10 matched modes are selected from amongst the original 20 modes of the cell lines; whereas the top 100 modes of the reference cell line (HSPC) are matched by up to 400 (original) modes of the other cell lines (**Figure 3.5B**).

The degree of *collectivity* of a given mode provides a measure of its distribution over different parts of the structure (Brüschweiler, 1995). Slower modes are usually more collective, cooperatively involving large groups of loci, and collectivity usually decreases with mode number, but this is not necessarily a smooth decrease. The collectivity of the top 500 modes for all chromosomes and cell types evaluated before and after matching the modes showed that the dependency of collectivity on mode number remained unaffected by mode-mode matching (**Figure 3.7**).

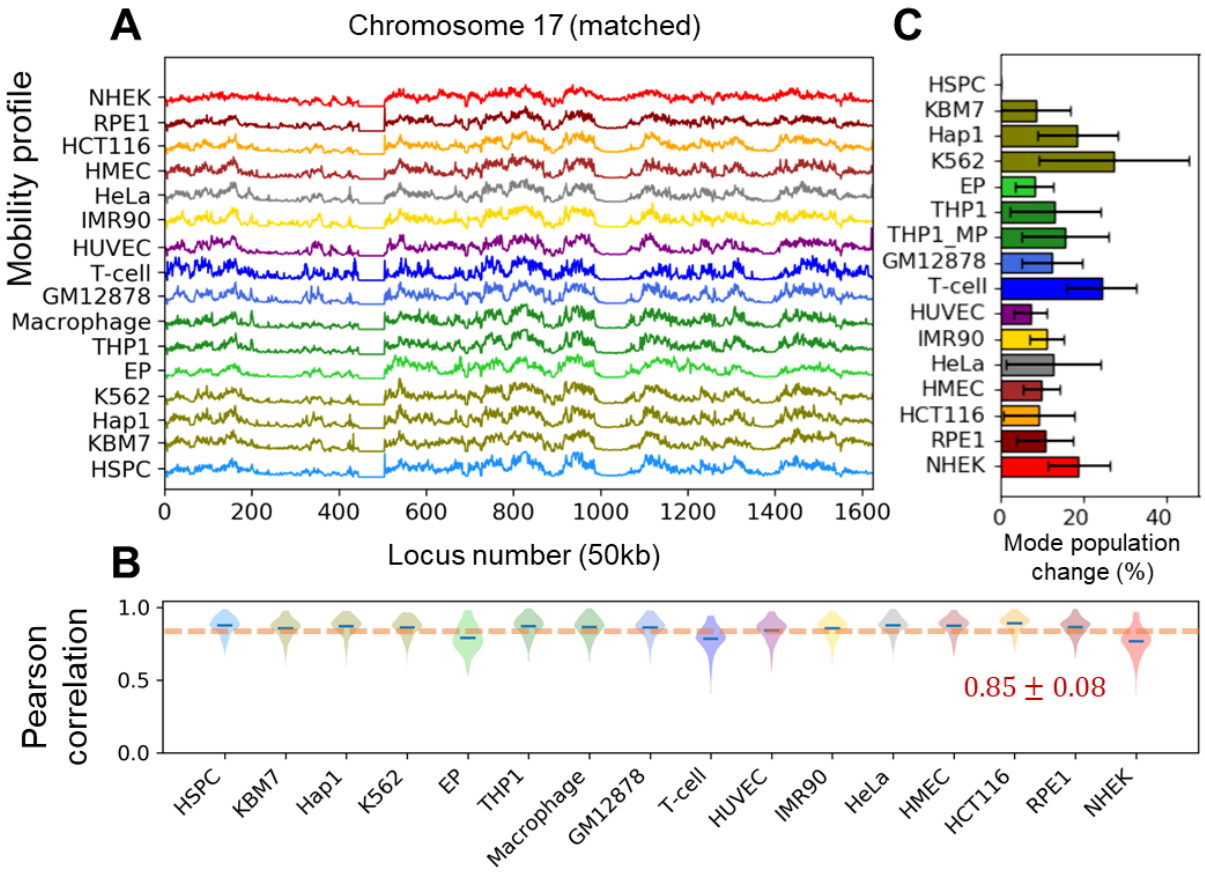

Figure 3.6 Verification of the close similarity of the spectrum of motions after eliminating the differences originating from the frequency dispersion. (A) Mobility profiles of genomic loci on chromosome 17 based on first 500 matched modes identified for all cell lines. **(B)** Distribution of Pearson correlations between the chromatin mobility profiles of each cell type and all others, obtained with the same set of modes. **(C)** Percent change in mode population after inclusion of equivalent modes for each cell line, averaged over all chromosomes. The *error bars* indicate the standard deviation among chromosomes. HSPC has no change because it is used as the reference. This figure is adapted from (Zhang, et al., 2020).

Overall, these results show that the chromatin of different types of cells have access to comparable modes of motion (encoded by their similar contact topologies) but not all of these pre-existing intrinsic modes are manifested, resulting in cell-specific mobilities of genomic loci. The differences in mobility profiles observed in **Figure 3.2** and **Figure 3.3** originate from the fact that the "active" modes differ between different cell lines. If we select the original softest 500 modes, we end up with cell-type specific mobility profiles. The profiles become similar only if we select the "equivalent" modes even though this set contains contributions from relatively "inactive" modes and excludes some "active modes".

In other words, *conserved* modes among different types of cells manifest themselves in a *signature profile* shared among all cell lines (**Figure 3.6A**) provided that the differences in the mode frequencies are suppressed. But in practice, not all modes are operative, and the fluctuations of genomic loci exhibit cell type-specific features. Some modes are "mute" while others are fully deployed, and which modes are selectively deployed depends on the cell type.

It is of interest to assess the fraction of the original modes that have been replaced by equivalent modes. Results are presented in **Figure 3.6C.** We note that despite showing the greatest enhancement to conform to the signature profile, NHEK is only in the third place to show largest mode changes (19%), indicating that some of the original slow modes for NHEK greatly differed from those for HSPC, and the substitution of those modes effectively restituted the mobility profile to closely resemble the signature profile. Two cell lines that showed the largest mode changes are K562 (27%) and T-cells (24%). Their mobility profiles exhibited an increase in average correlation with all others from 0.60 (each) to 0.86 and 0.78, respectively. EP experienced the least mode number changes (8.1 ± 4.6%), yet its average correlation increased significantly (from 0.54 to 0.79).

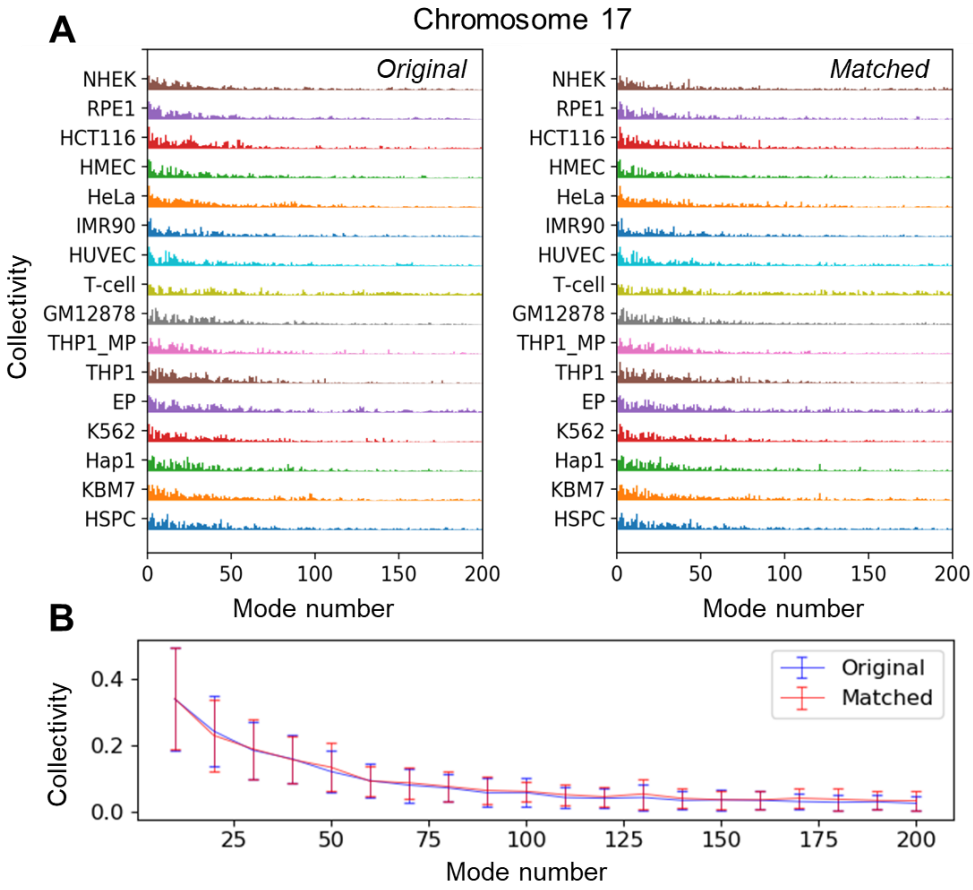

Figure 3.7 Collectivity profiles of GNM modes illustrated for chromosome 17 loci. (A) The *bars* display the degree of collectivity of each mode, in the range $1 \leq k \leq 200$, obtained for all cell types, before (*left*, original) and after matching the modes to those of the reference, HSPC (*right*, matched). **(B)** Mode collectivities averaged over all cells and plotted based on subsets of 10 modes demonstrate that the two sets display comparable distribution of collectivity in general. This figure is adapted from (Zhang, et al., 2020).

3.1.3 Discussion

The present comparative study of the intrinsic dynamics of chromosomes in a series of cell lines of different cell types using corresponding Hi-C data in the Gaussian Network Model (GNM) shed light to several fundamental features, including the shared fluctuation patterns in the spatial positions of different types of cells or *signature dynamics*, evident in the modes of motions in the

lowest frequency regime. These slowest (and most collective) modes of motion intrinsically accessible to the individual chromosomes of different types of cells were distinguished by their conservation, even among phenotypically divergent cells such as GM12878 and NHEK, yielding an average correlation cosine of $\langle S \rangle_1 = 0.84 \pm 0.18$ between all cell type pairs. The modes in the intermediate and high-frequency ranges, on the other hand, appeared to be less conserved. This is physically reasonable, because global modes, especially the first few, usually underlie the structural stability (Bahar, et al., 1998). The spatial organization of chromatin is hierarchical (see Section 2.2), and the conservation of global modes may suggest that the cells maintain similar upper levels of the hierarchy but organize the lower levels differently. This type of organization may ensure a stable genome structure and the framework to achieve cell type-specific gene transcription/regulation activities.

Careful analysis showed that the differences in the mode spectrum essentially resided in the contributions (statistical weights) of different modes of motions, rather than the availability of these modes of motions. In other words, different types of cells share pre-existing modes with similar "shapes" but different frequencies. While the ensemble of modes intrinsically accessible were comparable, some modes were "silent" in selected cell types, while others were "active". This distinction is reminiscent of the existence of the same set of genes in all cell lines but their differential expression levels in different cell types depending on the specific functions of these cells. Similarly, we have the same ensemble of collective motions theoretically accessible, but not all of them are operative within the same time window, and as a result the different types of cells end up exhibiting differential dynamics (**Figure 3.2A** and **Figure 3.3**). We demonstrated that the 16 cell types presently analyzed would have exhibited the same fluctuation behavior (**Figure 3.6A-B**), if their equivalent modes were equally active. This interesting finding brings important insights

into how genome structure may reorganize during cell differentiation and enable different accessibility patterns, shown here using a physical model at the genome-scale, for several types of cell lines.

3.1.4 Methods

Computational methods used here in this chapter are similar to the methods used in previous chapters with minor adjustments, such as the GNM analysis of the Hi-C data (Section 2.1.4.3), calculation of mode-mode overlap and spectral overlap (Section 1.2.4.5), and identification of equivalent modes (Section 1.2.4.3). Most of the calculations are performed using the application programming interface (API) ProDy (Bakan, et al., 2014; Bakan, et al., 2011). The specifications and adaptations for this study are described below.

3.1.4.1 Hi-C data acquisition and processing

The Hi-C datasets used in this study were downloaded from various sources (summarized in **Table 3.1**) using the Juicer/Straw tool (Durand, et al., 2016) implemented as an interface in ProDy (Bakan, et al., 2014; Bakan, et al., 2011). The full dataset contains cell types derived from different germ layers. The majority of the cell lines are hemopoietic cells of different types or at different developmental stages. Preprocessing steps are described in Section 2.1.4 and Section 2.1.4.2. GNM modes, MSFs and covariances of loci were evaluated for all 23 chromosomes in each dataset.

Linkage	Name	Germ layer	Reference
GM12878	Human lymphoblastoid cell line	mesoderm	(Rao, et al., 2014)
K562	Human immortalized myelogenous leukemia cell line	mesoderm	
KBM7	Chronic myelogenous leukemia (CML) cell line	mesoderm	
HUVEC	Human umbilical vein endothelial cell line	mesoderm	
IMR90	Lung fibroblasts	endoderm	
NHEK	Normal human epidermal keratinocytes	ectoderm	
HMEC	Primary mammary epithelial cells	ectoderm	
HeLa	Cervical cancer cells	ectoderm	
Hap1	Near-haploid human cell line derived from KBM7	mesoderm	(Sanborn, et al., 2015)
HCT116	Human colon cancer cell line	ectoderm	(Rao, et al., 2017)
RPE1	Retinal pigment epithelial cell line	ectoderm	(Darrow, et al., 2016)
HSPC	Hematopoietic stem and progenitor cells	mesoderm	(Joeng, et al., 2017)
EP	Erythroid progenitor cells	mesoderm	
T-cell	T lymphocytes	mesoderm	
THP1	Immortalized monocyte-like cell line	mesoderm	(Phanstiel, et al., 2017)
Macrophages	Macrophages derived from THP1	mesoderm	

Table 3.1 Dataset of cell lines analyzed in the present study. Cells that originate from hematopoietic stem cells are highlighted in *blue*. These all originate from mesodermal germ layers, as well as the human umbilical vein endothelial cell line (HUVEC, highlighted in *gray*). Those resulting from the differentiation of ectodermal germ layers are highlighted in *yellow*; and endodermal germ layers in *orange*. This table is adapted from (Zhang, et al., 2020).

3.1.4.2 Mode-mode overlaps

To compare two different sets of m modes defined as $\{(\lambda_i^A, \boldsymbol{v}_i^A) | i \in \mathbb{N} \text{ and } i < k\}$ and $\{(\lambda_j^B, \boldsymbol{v}_j^B) | j \in \mathbb{N} \text{ and } j < k\}$ obtained for the same chromosome of two different cell types A and B, for example, we evaluate the *mode-mode overlaps* organized in a correlation cosine map $S(A, B)$ the ij^{th} element of which is

$$[S(A,B)]_{ij} = \left| \boldsymbol{v}_i^{(A)} \cdot \boldsymbol{v}_j^{(B)} \right| \tag{3.1}$$

where $\boldsymbol{v}_i^{(A)}$ is the shape (eigenvector) of the i^{th} mode obtained for cell type A. $[S(A,B)]_{ij}$ varies in the range [0, 1], and the lower and upper limits refer to no and complete overlap, respectively.

Then, the level of conservation of mode i is evaluated by averaging $[S(A,B)]_{ii}$ over all pairs of (A, B), *i.e.*

$$\langle S\rangle_i = \frac{m(m-1)}{2}\sum_{A}\sum_{B,B\neq A}[S(A,B)]_{ii}\,, \tag{3.2}$$

where m is the total number of cell types (m = 16 here).

3.1.4.3 Identification of mode-mode-matches across different cell lines

Because of cell-type specific variations in the genome structure, the mode spectra also differ. Pairwise comparisons of the mode sets for different cell lines necessitate the identification of the *equivalent* (best matching) modes. As described, we first calculate the mode overlaps $[S(A,B)]_{ij} \in [0,1]$ for all eigenvector pairs (i, j) of cells A and B using equation 3.1, and then evaluate the cost of matching them as $1-[S(A,B)]_{ij}$ and finally select the mode pairs that minimizes the total cost using the Hungarian algorithm (Kuhn, 1955; Kuhn, 1956).

3.1.5 Acknowledgment

The presented work is part of the publication (Zhang, et al., 2020). I performed all the analyses. Dr. Ivet Bahar supervised the project.

3.2 Quantification of Differences in the Intrinsic Chromatin Dynamics Explains Cell Differentiation

3.2.1 Introduction

Numerous studies with biomolecular assemblies have demonstrated that accessibility to binding substrates does not necessarily map to functionality. A more important feature that enables function is the malleability of the putative active sites to optimize binding energetics and support adaptability to structural changes, manifested by conformational flexibility under physiological conditions (Haliloglu and Bahar, 2015). By analogy, it is reasonable to expect that genes located in loci distinguished by large amplitude fluctuations under equilibrium conditions would be more amenable to processing and expression. We performed a systematic comparative analysis to examine the existence of such correlations between the 3D mobilities of the genes and their expression levels. Using gene-set enrichment data based on RNA sequencing experiments deposited in Gene Expression Omnibus (GEO) (Barrett, et al., 2013), we demonstrated the existence of a strong coupling between cell-specific highly mobile genes (HMGs) predicted here by the GNM and the highly expressed genes (HEGs) compiled in the ARCHS4 database (Lachmann, et al., 2018).

Nonetheless, mobility profile of chromosomes is a 1D property and does not reflect the complexity of chromosomal dynamics. So, despite being shown to be related to important genomic properties such as chromatin accessibility and differential gene expression, locus mobility as a measure of cell type specificity would be incomplete because of the absence of inter-loci interactions. Such interactions between loci can be quantified by the cross-correlation map $\tilde{\mathbf{C}}$

derived from the GNM (see Section 2.1.4.3), which describes how much the pair of loci *i* and *j* are correlated with regard to their spatial movements, averaged over all possible modes of motions. Such correlations may originate from connectivity (sequence neighbors along the DNA), spatial proximity in the 3D genome or from "allosteric" effects involving other common connections. The mobility profile and directional cross-correlations thus provide complementary information on the respective sizes and orientational couplings of genomic loci movements.

3.2.2 Results

We first evaluated the GNM-predicted MSFs and cross-correlations for 23 chromosomes of 16 human cells (or cell lines) listed in **Table 3.1**. Then, we compared the GNM-predicted *highly mobile genes* (HMGs) with the *highly expressed genes* (HEGs) in multiple cell lines annotated in the ARCHS[4] database (Lachmann, et al., 2018). Finally, we computed the covariance overlap between the same chromosomes for every pair of cells to measure their similarities in terms of collective intrinsic dynamics.

3.2.2.1 Genes distinguished by high mobility correlate with those highly expressed in a cell-type-specific manner

We identified the subset of *highly mobile genes* (HMGs) distinguished by large amplitude motions (peaks in the mobility profiles, e.g. **Figure 3.2A**) in a given cell type but not in others, and explored the biological relevance of these strong departures from the average mobility profile of all cells (**Figure 3.14**), if any, to the differential function of the specific cells.

Specifically, we compared the HMGs predicted here with the *highly expressed genes* (HEGs) in multiple cell lines annotated in the ARCHS[4] database (Lachmann, et al., 2018) (see

Section 3.2.4.1). ARCHS[4] contains information on the HEGs of 125 common cell lines, obtained by integrating gene expression data from RNA-seq experiments deposited in the Gene Expression Omnibus (GEO) (Lachmann, et al., 2018). These 125 cell lines constitute our pool of "*candidate cell types*", which will be searched to explore the relationship between HMGs and HEGs. The pool contains HEG data for six of the 16 cell lines investigated here (K562, IMR90, HCT116, HUVEC, HeLa, and THP1), which will be referred to as the *query* cell lines. In each case, we screened the query cell line against the entire dataset of 125 candidate cell lines in ARCHS4 and computed the Jaccard index as a measure of the overlap between the HMGs of the query cell line and the HEGs of the candidate cell lines; and identified the top-ranking candidate cell lines whose HEG pattern shows the highest similarity to the HMG pattern of the query cell line. In each case we also display the results for the other 5 query cell lines for comparative purposes. Notably, the top-ranking candidate cell line turned out to be the query cell line itself in all cases (**Figure 3.8**).

Other top-ranking candidates also bear resemblances to the corresponding target as well. For instance, BJ cell (normal human foreskin fibroblast), NHDF (normal human dermal fibroblast) and MG63 (osteosarcoma with fibroblastic shape) all share a fibroblast-like morphology as IMR90, a fetal lung fibroblast. In the case of THP1, a typical cell model for primary monocytes, one of its top candidates, U937, also shows monocytic traits (**Figure 3.8**).

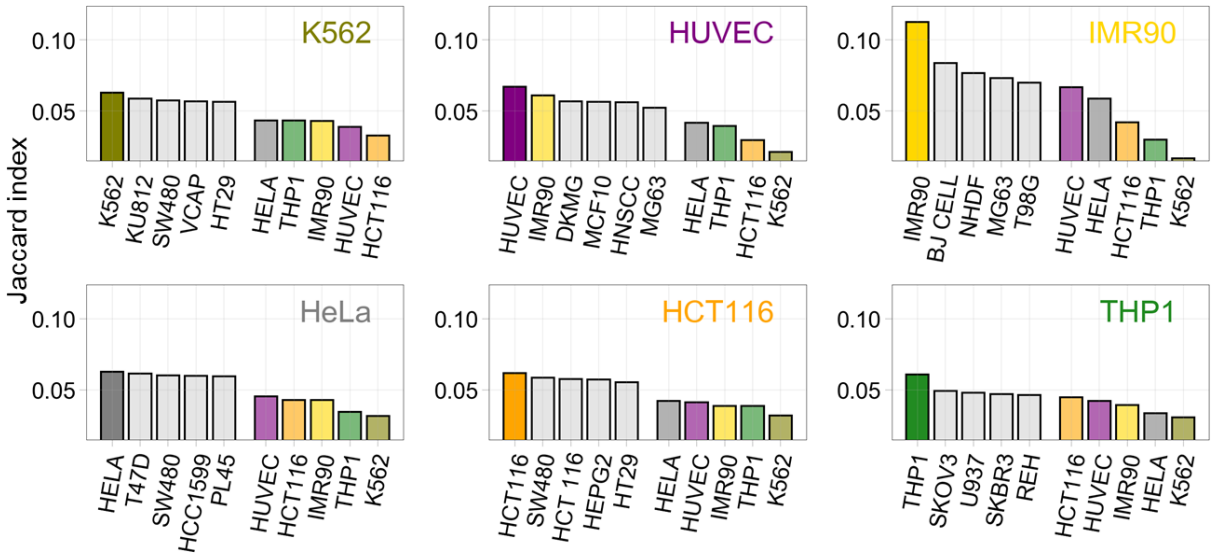

Figure 3.8 Overlap between cell-type-specific highly mobile and highly expressed genes. Results are presented for six of our dataset cell types that were represented in ARCSH[4]. The overlap was quantified using the Jaccard index. Top-ranking five (screened) cell lines whose HEGs exhibit the highest overlap with the HMGs of the query cell are shown by the *bar plots*. Overlaps between the HMGs of the query cell and the HEGs of other five cell lines are also presented for comparison. Strikingly, the top-ranking cell type (from the pool of 125 in ARCSH[4]) turns out to be the query cell type itself, demonstrating the distinctive overlap between HMGs and HEGs specific to each cell type. This figure is adapted from (Zhang, et al., 2020).

3.2.2.2 Locus-locus dynamical correlations show stronger dependency on cell type than do loci mobilities

As mentioned above, cross-correlations measure how much the movements of two loci are correlated. **Figure 3.9A** showed such cross-correlation maps for chromosome 17 as an example, computed for all cell lines in our dataset. We observed strong correlations among sequential neighbors (*red band* along the diagonal). Pronounced couplings are observed within selected regions presumably representing TADs (*red squares* on the diagonal). As to the cross-correlations between sequentially distant loci, a range of behavior is detected. For instance, an interesting

pattern near the end of the long arm (loci 1,000-1,500 approximately) is distinguished in K562: two distal domains exhibit a pronounced coupling (highlighted by the *black square* in **Figure 3.9A** and schematically depicted in **Figure 3.9B**). This behavior tends to be more prominent in mesodermal cell lines (including hematopoietic cell lines labelled in *blue* and HUVEC labelled in *gray*), than in ectodermal cell lines such as NHEK, HMEC, HCT116, and RPE1. We also notice that, as compared to other cell lines, K562, HCT116, THP1, and macrophage exhibit stronger cross-correlations among the loci in the short arm (loci 1-500), implying a higher packing density in the region. Another manifestation of tight packing is the suppressed mobility of loci occupying such regions, noted earlier in the short arm region (**Figure 3.2A**). Thus, tightly packed regions which exhibit minimal movements are also distinguished by their close directional couplings, in accord with their restricted movements as almost rigid blocks. This feature can be observed clearly by displaying the MSF profiles along the axes of the cross-correlation maps. The *black arrows* in **Figure 3.10** highlight such regions.

We then examined the overlaps between covariance matrices obtained for different cell lines (see Section 3.2.4.2). Unlike fluctuation profiles, covariance matrices showed higher variations among the cells. The overall covariance overlap averaged over all chromosomes and pairs of cell lines was 0.48 ± 0.11 (*blue violins* in **Figure 3.9C**). NHEK again yielded the lowest average overlap of 0.40 ± 0.10, however, it was not an outlier, and many other cells exhibited comparable values. The overlaps between the covariance matrices could be slightly improved upon mode matching (*red violins* in **Figure 3.9C**), but the improvement was much more limited compared to that observed in locus mobility profiles. Overall, this analysis the couplings between loci movements exhibit a stronger dependency on cell type than that the mobility profiles of individual loci presented in Section 3.1.2.1.

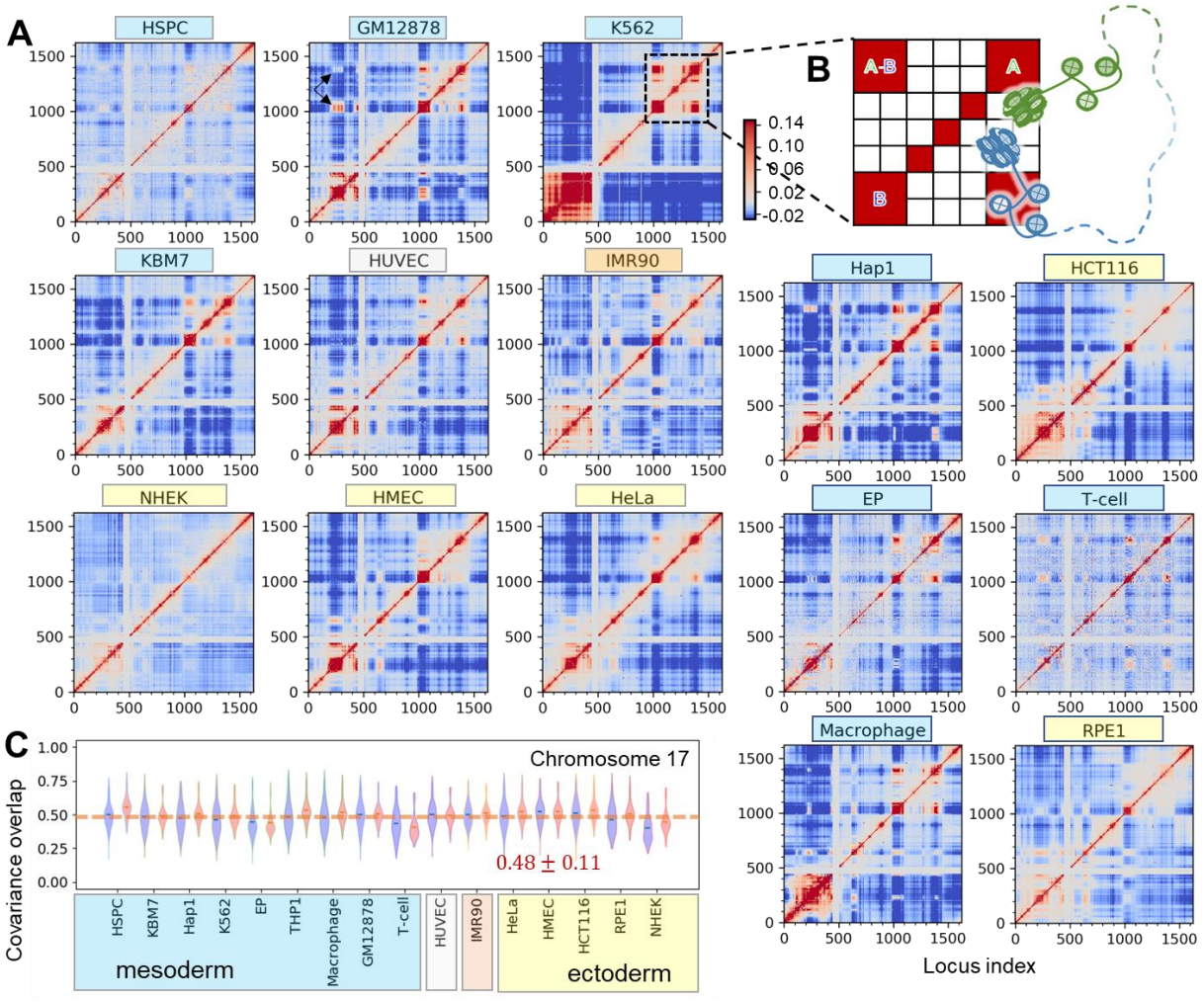

Figure 3.9 Locus-locus cross-correlations reflect cell-type specificity. (A) Cross-correlations $\tilde{\mathbf{C}}_{ij}$ between genomic loci computed for chromosome 17, shown for 15 different cell types. $\tilde{\mathbf{C}}_{ij}$ values vary from -0.26 (anticorrelated, off-diagonal regions in *dark blue*) to 1.0 (fully correlated, diagonal elements, in *red*). The white bands at loci ~400-500 refer to the centromere, where Hi-C contact data are missing. **(B)** Schematic description of chromosomal organization indicated by the correlation patterns. *Red blocks* A and B on the diagonal represent two domains (A: *green*; and B: *blue*) with tightly packed DNA; and the off-diagonal red blocks indicate the long-range domain-domain couplings between A and B. The *dashed curve* depicts a long sequence not shown between A and B. **(C)** Covariance overlaps among cell lines averaged over all chromosomes based on the first 500 original (*blue violins*) or matched (*red violins*) modes. This figure is adapted from (Zhang, et al., 2020).

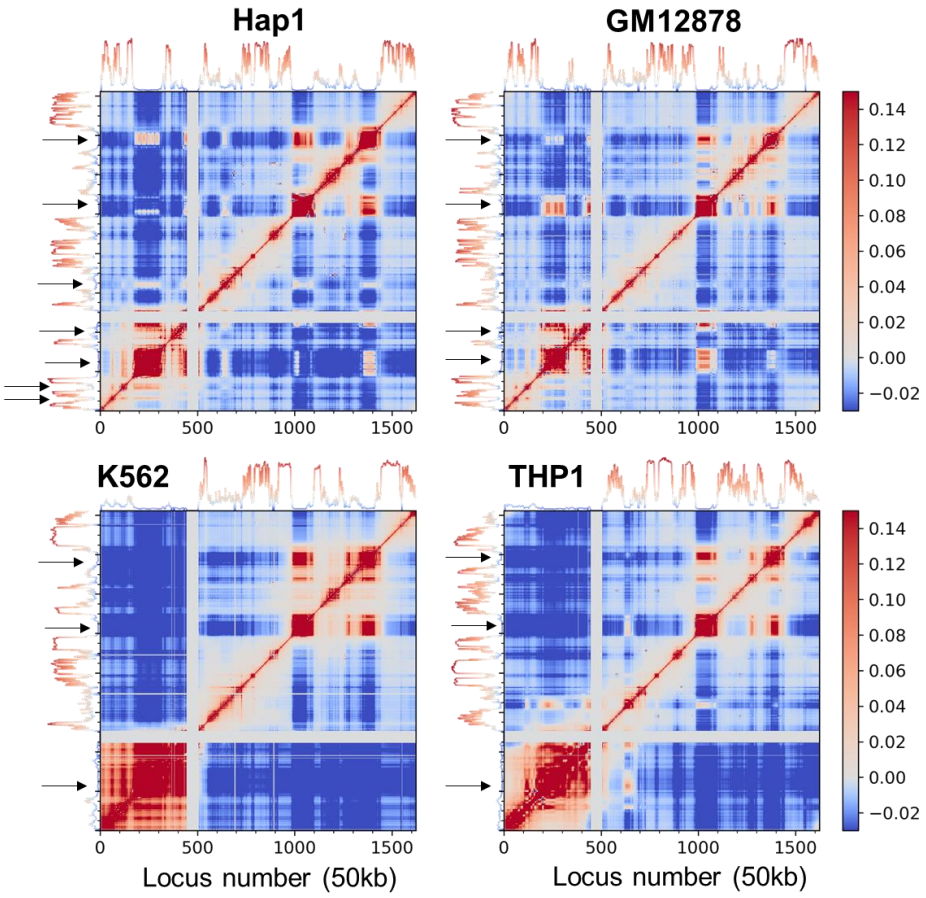

Figure 3.10 Cross-correlation maps aligned with MSFs illustrated for four cell types. The heat maps display the correlation cosines between the movements of chromosome 17 loci of the different types of cells (*labeled*), based on 500 softest modes. *Black arrows* along the *left ordinate* show examples of regions that exhibit high directional correlations while their mobility is low, indicative of severe spatial restrictions (minima in mobility profiles) constraining the loci to move together, or to be rigidly held together while undergoing small fluctuations. This figure is adapted from (Zhang, et al., 2020).

3.2.2.3 Covariance overlap between loci as a discriminative metric for assessing the divergence of cell lines

To understand the impact of cell-type-specific locus-locus dynamical couplings on the response or adaptation of cells to endogenous or environmental effects, on cell differentiation, we

quantified the differences between the covariances obtained for individual chromosomes in different cell lines using as metric the covariance overlap (see Section 3.2.4.2) and performed a series of experiments *in silico*.

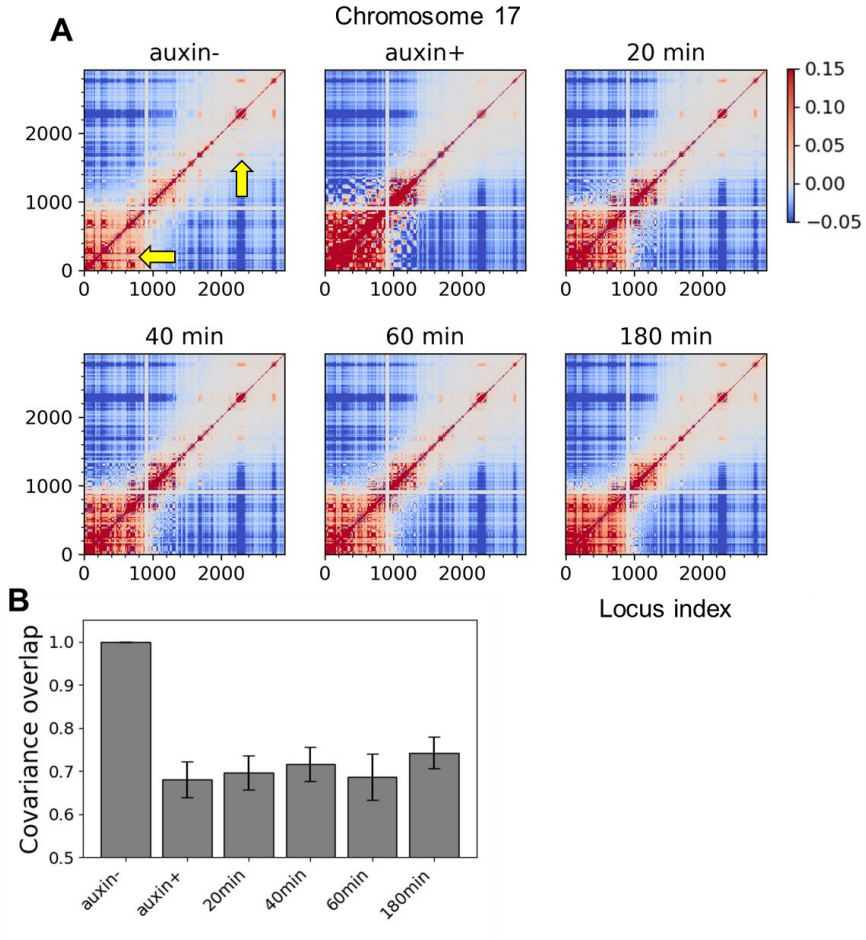

Figure 3.11 Time evolution of chromatin contact topology after auxin treatment. (A) Covariance matrices for chromosome 17 of HCT116 cells before and after auxin treatment, and 20, 40, 60, 180 minutes after auxin has been withdrawn. It can be seen, indicated by yellow arrows, loci interactions are greatly weakened or disrupted after auxin treatment (compare the matrix for auxin+ with that for auxin-). These interactions are gradually restored during time after withdrawn (compare the matrices for 20, 40, 60 and 180 minutes). Hi-C data are obtained from (Rao, et al., 2017). **(B)** Average overlaps between the covariance matrices computed based on Hi-C maps at each time point after the treatment of auxin and those computed based on Hi-C maps for normal HCT116 (Rao, et al., 2017). Standard deviations are shown by *error bars*. This figure is adapted from (Zhang, et al., 2020).

First, we examined the loop domain loss for HCT116 under the influence of auxin using time-dependent Hi-C dataset, mainly Hi-C maps for HCT116 cells under normal conditions (auxin-), 6 hours after the treatment of auxin (auxin+), and 20, 40, 60 or 180 minutes of auxin withdrawal (Rao, et al., 2017). We evaluated the covariance overlaps between the covariance (of all chromosomes) of the treated cells at each time point and those of the normal cells. As expected, the average covariance overlap dropped by approximately 30% after the treatment and gradually recovered with time after auxin withdrawal (**Figure 3.11**).

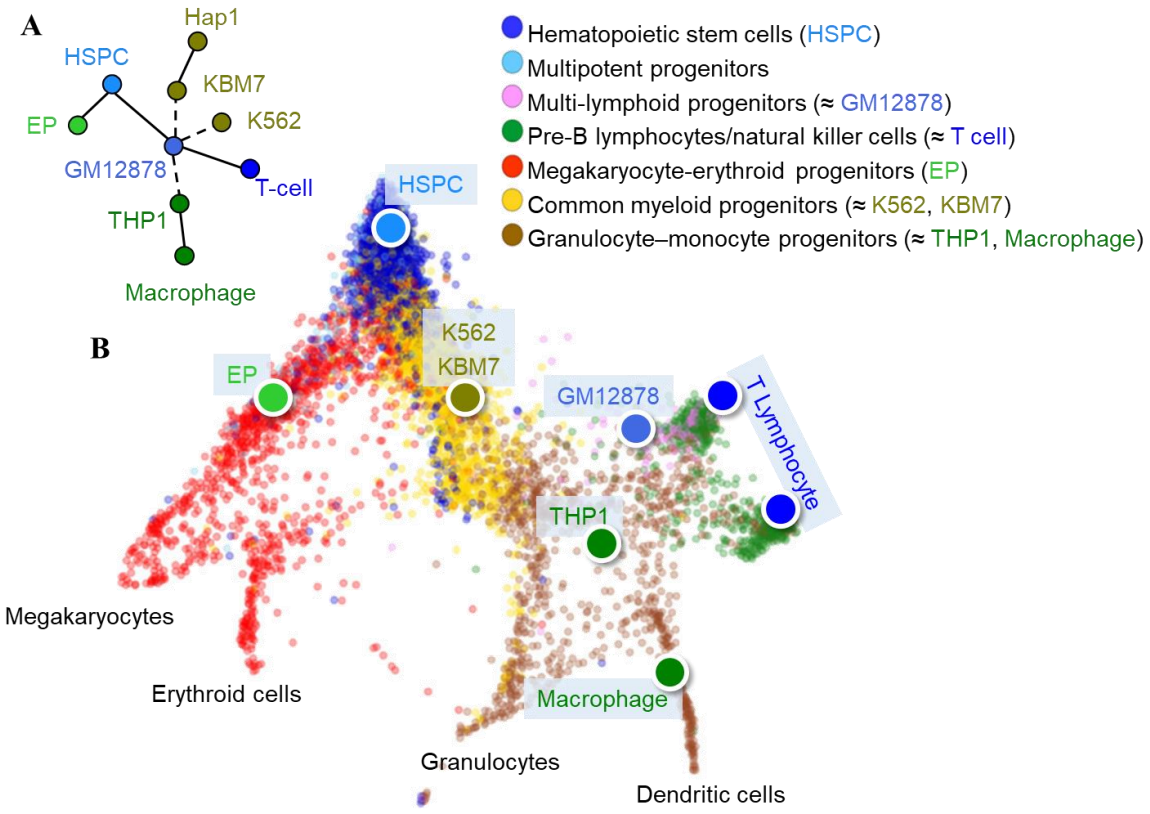

Figure 3.12 Hematopoetic cell relationships represented by a tree determined by their differentiated chromatin dynamics. (A) Collective MST for hematopoietic cells based on the covariance overlaps computed for all chromosomes. **(B)** *k*-nearest neighbor-based clustering/visualization of single-cell RNA-seq data of hemopoietic progenitors, adapted from (Pellin, et al., 2019). Cell lines that used in this study are marked by *nodes* at corresponding (approximate) positions on this map and *color coded* consistently with panel A. This figure is adapted from (Zhang, et al., 2020).

Second, we asked whether the differences in covariances could be used to distinguish cell types. To answer this question, we constructed a distance graph for the cell lines based on the covariance overlaps, where each node represents a cell line and each edge is weighted by the arc distance d_{cov} between the covariance matrices, obtained for the corresponding pair of cells (see equation 3.7). We then determined the MST that revealed the relations between cell lines based on their covariance. We applied this procedure to the hematopoietic cell lines because among the cell lines we collected those had the clearest differentiation hierarchy and lineages in earlier or intermediate stages, such as HSPC, GM12878, and THP1. Covariance overlaps obtained for different chromosomes gave rise to different MSTs (data not shown), possibly due to different rates in the spatial organization of different chromosomes during cell differentiation. To determine the MST for all chromosomes, we constructed a graph based on the maximum $d_{cov}(A, B)$ (see Section 3.2.4.3) between all cell types, which led to a collective MST that retain only the closest relationships among cell types in terms of genome structure similarities. The resulting MST was found to be broadly agree with single-cell transcriptional behavior of hematopoietic progenitors (Hay, et al., 2018; Pellin, et al., 2019) (**Figure 3.12**). The tree correctly reproduces the transcriptional similarities among blood cell lineages (indicated by *solid edges*), including the fact that Hap1 is derived from KBM7. K562 and KBM7, both of which relate to myeloid progenitors, should have been closer to HSPC than GM12878, a lymphoblastoid cell line. This discrepancy might originate from the cancerous nature of K562 and KBM7 (hence the *dashed edges* between KBM7, K562 and GM12878). Moreover, the relationship among monocyte progenitors (THP1) and GM12878 is ambiguous, also marked by a *dashed edge* in **Figure 3.12A**.

Third, we applied the neighbor-joining method to construct a "phylogenetic" tree based on the maximum covariance distance map obtained for all investigated cell lines. The resulting tree

groups together similar cell lines, e.g. hematopoietic cells cluster together except for HSPC and EP, and epithelial cell lines, NHEK, RPE1, and HCT116, are under the same branch (**Figure 3.13**). Interestingly, one of the two leukemia cell lines, K562, is clustered with HeLa derived from the cervical tumor, whereas the other, KBM7 and its derivative Hap1, are grouped with normal lymphoid cells, GM12878 and T-cells, suggesting cancer heterogeneity among leukemia cells.

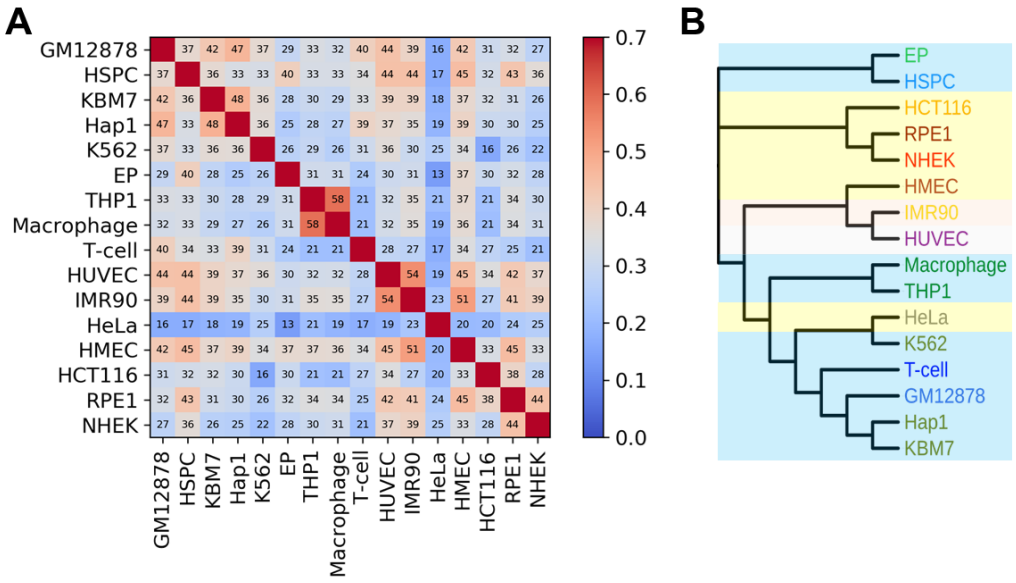

Figure 3.13 Collective measurement of similarities between different cell types in terms of their chromatin dynamics. (A) Minimum covariance overlaps (correspond to maximum covariance distance) across all chromosomes obtained for all cell pairs. The numbers show the overlap as percentages. **(B)** Neighbor-joining tree for all cell lines constructed using the minimum (maximum) covariance overlaps (distances) for all chromosomes. Branch lengths are not proportional to arc distances. The *color shades* are added to facilitate the visualization of the grouping of ectodermal, mesodermal and other cell lines as in **Figure 3.9C**. This figure is adapted from (Zhang, et al., 2020).

The similarities between cell lines found here at the chromatin dynamics level are in accordance with an earlier study (Sauerwald and Kingsford, 2018) where they found that HMEC, despite being an epithelial cell line originated from the ectoderm, was more similar to endodermal and mesodermal cells; the leukemia cell lines, KBM7 and K562, resemble GM12878; and there

are similarities between IMR90 and HUVEC. However, the results for epidermal keratinocytes, NHEK, are different. High similarities between the TADs identified for NHEK and those for GM12878 and K562 were reported (Sauerwald and Kingsford, 2018), while NHEK shows little resemblance to other cell lines in terms of its intrinsic dynamics.

3.2.3 Discussion

As discussed in Section 3.1, chromatin accessibility plays an essential role in regulating gene expression and cell differentiation by allowing or preventing physical interactions between transcription factors or other regulatory proteins and genomic loci (Klemm, et al., 2019). Theoretically, the accessibility of a site is predominantly determined by the local packing density, which is manifested by high mobilities in the 3D fluctuation profiles predicted by the GNM (Section 2.1.2.1). Then, to understand the role of high mobility at selected loci in defining cell differentiation, we identified cell type-specific variations in the equilibrium dynamics that give rise to genes that are specifically more exposed/mobile in one cell than another, i.e. HMGs. A distinctive overlap between HMGs and HEGs has been found upon a systematic examination with the ARCSH[4] database. The analysis demonstrates that (i) the unique HMG pattern predicted here to typify each cell line strongly correlates with the cell-line-specific HEG behavior, suggesting a strong link between high mobility and high expression, and (ii) the set of HMGs provides a sufficiently distinctive feature to accurately discriminate between cell lines exhibiting different expression patterns. It also suggests that (iii) high conformational flexibility or spatial mobility may be a prerequisite for enabling productive interaction with proteins and thereby effective transcription or gene expression.

Cross-correlation maps of chromosomal loci is an additional and maybe more important feature that provide additional and important information on locus-locus spatial couplings/correlations which contribute to the chromatin dynamics in a three-dimensional (3D) space. Here we compared the covariance matrices using the covariance overlap, which is a well-established metric to compare the subspaces spanned by normal modes and has been used in many applications (Grossfield, et al., 2007; Hess, 2002; Romo and Grossfield, 2011). We found that while the locus mobilities shared much resemblance among cell lines, long-range couplings measured by spatial cross-correlations (covariance) exhibited more diversities, even among closely related cells (**Figure 5A** and **Supplementary Figure S5A**). For example, for chromosome 17, the off-diagonal cross-correlations are much weaker and sparser for K562 than for other hematopoietic cells (**Figure 5A** and **Supplementary Figure S5A**), which may suggest that in K562 the two arms of chromosome 17 are partially disordered, if not unfolded. Moreover, in the same chromosome, there are two anchor regions (**Figure 5A**, *black arrows*) that connect the two arms of the chromosome in GM12878; whereas the regions are absent in two leukemia cell lines, K562 and KBM7, as well as in THP1 and all four epithelial cell lines. These observations agree with the view that while chromatin domain positions in space are stable during differentiation, interactions within and between domains can change drastically (Dixon, et al., 2015). Furthermore, the cell trees based on covariance overlaps did capture some lineage relationships and further suggested the utility of pairwise covariance overlaps as a metric for quantifying the differentiation of cells with regard to their collective dynamics.

Overall, this analysis revealed strong overlap between highly expressed genes and those distinguished by high mobilities; and cross-correlations of genomic loci are cell type-specific. These observations, with important implications in cell differentiation, invites attention to the

significance of the intrinsic mobilities of the individual genes in enabling their transcriptional regulation, shown here using a physical model at the genome-scale, for several types of cell lines.

3.2.4 Methods

3.2.4.1 Overlap between HMGs and HEGs

Relative mobilities of genomic loci are calculated by subtracting from the MSF profile $\langle \Delta \boldsymbol{r}_i^2 \rangle^{(A)}$ of locus *i* in cell type *A* the average over all *m* cell types, i.e.

$$\Delta\langle \Delta \boldsymbol{r}_i^2 \rangle^{(A)} = \langle \Delta \boldsymbol{r}_i^2 \rangle^{(A)} - \frac{\sum_A \langle \Delta \boldsymbol{r}_i^2 \rangle^{(A)}}{m}. \tag{3.3}$$

Loci with the highest $\Delta\langle \Delta \boldsymbol{r}_i^2 \rangle^{(A)}$ (top 10%) are considered as highly mobile, and genes within these loci are called "highly mobile genes" (HMGs) for that cell type *A*. The ARCHS4 database (Lachmann, et al., 2018) from Enrichr (Chen, et al., 2013; Kuleshov, et al., 2016) contains HEGs data for 125 cell types. We used Jaccard index as a metric to evaluate the overlap between HMGs for cell line *A* and the HEGs for cell line *B* from the ARCHS4 database,

$$J\left(HMG^{(A)}, HEG^{(B)}\right) = \frac{\left| HMG^{(A)} \cap HEG^{(B)} \right|}{\left| HMG^{(A)} \cup HEG^{(B)} \right|}. \tag{3.4}$$

Then, the following computational test/protocol of four steps, schematically described in **Figure 3.14A** is adopted: (1) we compute for a given cell type *A* the *relative mobilities* of genomic loci with respect to the average over all cells examined by the GNM, (2) we identify those loci, or corresponding genes, which rank in the top 10% in terms of their mobility. These are the HEGs specific to cell type *A*. (3) The list of HEGs is provided as input to search the pool of candidate cell types and extract those cell types *B* whose HMGs provide maximal overlap (highest Jaccard indices) with the HEGs of cell type *A*. (4) the top-ranking 5-6 candidate cells resulting from this

screening process are shown in the *bar plots* in **Figure 3.8** (*left bars*), along with the results for the other GNM-characterized cell types also contained in the pool (color-coded; *right bars*).

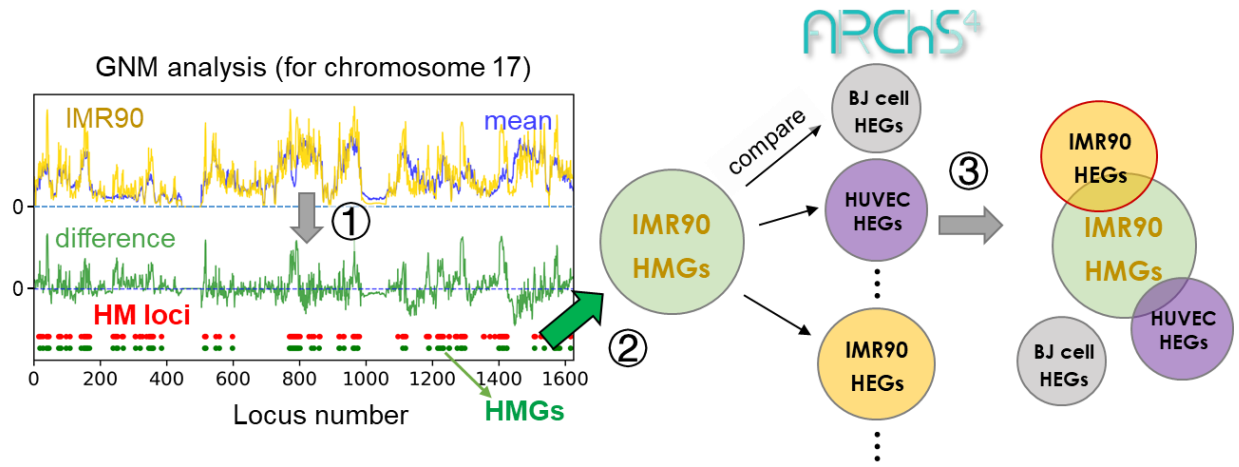

Figure 3.14 Illustration of the 4-step protocol in silico test of the relationship between HMGs and HEGs. For IMR90 as a query cell: **(1)** Identification of high mobility (HM) loci from the difference (*green curve*) between the mobility profile obtained with the GNM (shown here for IMR90 chromosome 17; *yellow curve*), and that averaged across all 16 cell lines (*blue curve*); HM loci are defined as those exhibiting the top 10% mobility, shown by the *red dots*. **(2)** Genes located in HM loci, HMGs, are identified (shown by *green dots*). The procedure is repeated for all chromosomes, and the resulting list of HMGs for IMR90 chromatin is used for screening against the cell-type specific HEGs compiled in the ARCSH[4] database; **(3)** Similarities between the HMGs of the query cell type and the HEGs of the 125 cell types in ARCSH[4] are measured by the Jaccard index, and rank-ordered for each query cell type. This figure is adapted from (Zhang, et al., 2020).

3.2.4.2 Covariance overlap for quantifying the similarities of chromatin dynamics

The similarities between covariance matrices $\mathbf{C}_A$ and $\mathbf{C}_B$ for respective cell types A and B is quantified by the spectral/covariance overlap (Hess, 2002) (as described in Section 1.2.4.5):

$$SO(A,B) = 1 - \left[\frac{\sum_{i=1}^{n-1}(\sigma_i^A + \sigma_i^B) - 2\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}(\sigma_i^A\,\sigma_j^B)^{\frac{1}{2}}\,(\boldsymbol{v}_i^A \cdot \boldsymbol{v}_j^B)^2}{\sum_{i=1}^{n-1}(\sigma_i^A + \sigma_i^B)}\right]^{\frac{1}{2}}. \qquad (3.5)$$

Here n is the total number of nodes (meaning n – 1 non-zero modes in total). σ_i denotes the variance of mode i, equal to the reciprocal of λ_i. An important difference from the procedure described in Section 1.2.4.5 is that, because Hi-C maps are measured for different cell lines that may have different total read counts, we normalized the variances by

$$w_i = \frac{\sigma_i}{\sum_{j=1}^{n-1} \sigma_j} = \frac{1/\lambda_i}{\sum_{j=1}^{n-1} 1/\lambda_j}. \tag{3.6}$$

w_i serves as a *prior probability* of contribution from mode i. This normalization permits us to directly compare the covariance matrices derived from different datasets.

3.2.4.3 Cell dendrograms based on chromatin dynamics

The spectral overlap can be converted to an arc distance (spectral distance) as

$$d_{cov}(A, B) = \arccos\big(SO(A, B)\big). \tag{3.7}$$

We took the maximum spectral distances across all chromosomes for each cell pair (*A*, *B*) to construct a distance graph G_D where the vertices represent the cells, and the edges are weighted by the spectral distances between the corresponding vertices. For characterizing the cellular hierarchy among hematopoietic cells, a minimum-spanning tree (MST) was found using Prim's algorithm (Prim, 1957). This way, cell lines at intermediate stages are treated as internal nodes. For all cells, because of the absence of intermediate cell lines, neighbor-joining (NJ) algorithm (Saitou and Nei, 1987) was adopted, where all cell lines are treated as terminal nodes.

3.2.5 Acknowledgment

The presented work is part of the publication (Zhang, et al., 2020). Fangyuan Chen contributed to the comparison between HMGs and HEGs. I performed the rest of the analyses. Dr. Ivet Bahar supervised the project.

Future Directions

Since their introduction in the late 1990s, the ENMs have become a widely used tool for exploring the unique functional motions of protein molecules over the years. The versatility and flexibility of the ENMs lend themselves to not only convenient and efficient applications to biomolecular systems, but also many methodological extensions that complement and expand the original models or the existing computational toolset (Atilgan and Atilgan, 2009; Eyal and Bahar, 2008; Hinsen, et al., 2000; Kaynak, et al., 2018; Kurkcuoglu, et al., 2016; Lezon and Bahar, 2012). In this work, we focused on developing computational frameworks on top of ENMs to explore/enable their capability in processing large datasets. Our study demonstrated two novel applications of the ENMs, the first on the analysis of protein family dynamics and the second on the evaluation of the chromatin intrinsic dynamics using information on coarse-grained contact topology. Below, we will briefly recap the conclusions reached from the analyses described in each chapter and discuss future directions of improvements or applications.

In Chapter 1.0, we developed and used an integrated computational pipeline, SignDy, for retrieving and analyzing the signature dynamics of protein families. We found that family members share conserved global modes that presumably provide the architecture for the fold to perform its main molecular functions, and motions in the LTIF regime define the specificity of subfamilies. In the future, additional studies with SignDy by a wide range of users with expertise in particular proteins and families would provide deeper insights into the evolution of dynamics and its importance for function. A reasonable strategy for utilizing SignDy in characterizing family/subfamily dynamics *vis-à-vis* structure and function evolution would be: (i) generate the

mode conservation and collectivity profiles for the investigated family; (ii) identify the conserved modes in different regimes; (iii) examine the corresponding mode shapes to (iv) identify critical sites responsible for the evolutionarily conserved signature dynamics (minima in global modes) and stability (peaks in HF modes) as well as those susceptible to subfamily-specific divergence (in conserved LTIF modes); and (v) generate dendrograms that provide information on dynamics similarities in different regimes, complementing sequence and structure similarities, among family members. While subfamily-subfamily spectral distances have been analyzed in Section 1.2.2.6 based on different frequency windows of structural dynamics (**Figure 1.14**), computations may be performed for narrower windows or even individual modes, to identify the most discriminative modes and infer new design/engineering principles for alterations of function.

In Chapter 2.0, we adapted the GNM to modeling chromatin's intrinsically accessible dynamics, and we made predictions on several dynamic properties of the chromatin molecule, which were shown to satisfactorily compare with data from chromatin accessibility (DNase-seq and ATAC-seq) and ChIA-PET experiments, and known chromatin substructures such as compartments and TADs. Future GNM analyses of chromatin dynamics could focus on the nature of the long-range couplings, analysis of their biological significance, or the meaning of genomic regions that exhibit high covariances. GNM also predicts a measure of the overall coupling of each genomic locus to others (i.e. the covariance matrix), the significance of which requires further investigation. The GNM was shown to capture several biological properties of chromosomes, but further insights into cooperative events, including the inter-chromosomal interactions, is within reach by focusing on the softest (lowest frequency) modes of motion predicted by the GNM. Modeling such inter-chromosomal interactions could be further exploited by a systems-environment framework (Hinsen, et al., 2000) to help ameliorate the tip effect at certain

chromosomal regions induced by sparse contacts. Finally, gene transcription is believed to be partly regulated by chemical modifications to histone proteins, and it has been shown in previous studies that histone modification markers (HMMs) are strongly associated with domain boundaries (Filippova, et al., 2014). It would be interesting to perform an analysis similar to what we did in Section 2.2.2.4 with the APBSs to investigate if/how HMMs are differentially enriched at domain boundaries, and how this enrichment depends on the hierarchical depths. In addition, several potential directions for the application of HiDeF to other types of biological networks were discussed in Section 2.2.3.

In Chapter 3.0, we showed that the comparative analysis of the fluctuation spectrum and CCDDs can reveal the differences across cell types. Similar to the case of protein families examined in Chapter 1.0 where global dynamics were found to be conserved, it would be of interest to explore whether cell-cell variabilities as well as the differences in disease *vs* normal states could equally be rationalized in the light of chromatin dynamics as more data become accessible on cell type-specific genome spatial organization. As more data will become available, more detailed analytical treatments using broader datasets, including more extensive single cell Hi-C data, will help obtain more complete and accurate information on cell-specific chromatin dynamics as well as their relevance to cell differentiation. Last but not least, a comparison of the hierarchies of chromatin domains across different cell types/species may provide more insights into the cell type-specific chromatin dynamics, to enhance our understanding of what type (extent) of structural rearrangements could lead to the observed changes in frequency dispersion while retaining similar mode shapes.

Appendix A Relationship between Spectral Clustering and the GNM

Spectral clustering. In early studies (Chan, et al., 1994; Hagen and Kahng, 1992), spectral clustering (or clustering based on the *ratio cuts* criterion) has been shown to relate to the *weighted quadratic placement problem* (QPP), which aims to find the optimal locations of n points $\boldsymbol{x} = [x_1, x_2, \dots, x_n]$ to minimize their total weighted squared distance:

$$z = \sum_{ij} a_{ij}\left(x_i - x_j\right)^2. \tag{A. 1}$$

The non-negative weights a_{ij} can be organized into an $n \times n$ affinity matrix, i.e. $\mathbf{A} = \left[a_{ij}\right]$, essentially defining a network topology of $\boldsymbol{x}$. Suppose $\mathbf{D}$ is the degree matrix of the network, and $\mathbf{L} = \mathbf{D} - \mathbf{A}$, the QPP can be expressed as a quadratic programming problem:

$$\text{minimize}_{\boldsymbol{x}}\ z = \boldsymbol{x}^T\mathbf{L}\boldsymbol{x}, \qquad \text{subject to } \boldsymbol{x}^T\boldsymbol{x} = 1. \tag{A. 2}$$

The constraint is to avoid the meaningless solution that $\boldsymbol{x}$ are all zeros. By introducing the Lagrange multiplier and setting the derivative to zero, it arrives at the following eigenequation:

$$\mathbf{L}\boldsymbol{x} = \lambda^{\dagger}\boldsymbol{x}. \tag{A. 3}$$

It can be seen that the solution to the QPP is the eigensystem of $\mathbf{L}$, the Laplacian matrix that defines the network topology. The first k eigenvectors $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_k]$ are the optimal placement of the points in k dimensions (i.e. Laplacian embedding), and the sum of the first k eigenvalues $\lambda_1^{\dagger}, \lambda_2^{\dagger}, \dots, \lambda_k^{\dagger}$ is the cost of the placement and total squared distance among the points. Laplacian embedding serves as (one of) the justification of spectral clustering and it will be discussed below its connections to the GNM.

An optimization view of the GNM. Minimizing the GNM potential (equation 2.3) under the constraint that the atomic displacement $\Delta \boldsymbol{r}$ is normalized leads to the following optimization problem:

$$\text{minimize}_{\Delta \boldsymbol{r}}\, V_{GNM} = \frac{1}{2}\Delta \boldsymbol{r}^T \boldsymbol{\Gamma} \Delta \boldsymbol{r}, \qquad \text{subject to } \Delta \boldsymbol{r}^T \Delta \boldsymbol{r} = 1,$$

which similarly to equation A. 2, arrives at an eigenvalue problem:

$$\boldsymbol{\Gamma} \Delta \boldsymbol{r} = \lambda \Delta \boldsymbol{r}, \tag{A. 4}$$

$\boldsymbol{\Gamma}$ is the Kirchhoff matrix which takes the form of a Laplacian. Eigenvectors and eigenvalues of $\boldsymbol{\Gamma}$ correspond to shapes and frequencies of normal modes. Equation A. 3 and A. 4 connect the GNM to spectral clustering in that they both solve the problem by retrieving the eigenvalues of a Laplacian matrix. Notably, the zero eigenvector corresponds to the translation of the entire network as a rigid body in the GNM; and in QPP, this correspond to a trivial solution where every point is placed at the same location, however, it was shown in (Stella and Shi, 2003) that this seemingly trivial zero eigenvector is just as important as any others for generating the optimal clustering solution.

Despite their similar mathematical forms, the physical interpretations of the two methods are distinct. Specifically, the variables $\boldsymbol{x}$ in the QPP are spatial positions of the points/nodes whereas in the GNM $\Delta \boldsymbol{r}$ are the *changes* in the positions. This means that while spectral clustering seeks to cluster nodes based on their proximity in an embedding space, the GNM reveals and identifies nodes with similar dynamics. Nonetheless, this difference does not compromise the utility of applying spectral clustering techniques on the GNM results, but sheds light on the fact that the "dynamically coupled" domains identified by the GNM should be also physically close, if a global distance optimum is reached.

Appendix B Inferring Hierarchy from Multiresolution Clustering Results

Background. Given n data points, $x_1, x_2, \ldots, x_n$, let **P** be an $n \times m$ binary matrix that denotes m non-exclusive clusters identified by some algorithm(s) using different parameter settings. Each column of **P** is a binary vector that indicates which data points are owned by which cluster, $\mathbf{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_m] = [p_{ij}]$,

$$p_{ij} = \begin{cases} 1 & \text{if } x_i \text{ is owned by cluster } j, \\ 0 & \text{otherwise.} \end{cases} \quad \text{(B. 1)}$$

Pairs of clusters may be *disjoint*, *overlapping* with, or *containing* each other. These relationships will be identified and represented by a *directed acyclic graph* (DAG) in later steps.

Containment index. Given two sets (clusters) of data points, A and B, we define *containment index* of *B* with respect to *A* as follows, to measure how much of B is included in A:

$$ci(A, B) = \frac{|A \cap B|}{|B|}. \quad \text{(B. 2)}$$

Obviously, $ci(A, B) \in [0,1]$. The metric is asymmetric provided that A is different from B. If $ci(A, B) > \kappa$ and $ci(A, B) > ci(B, A)$ with $\kappa \in (0.5, 1.0]$, it is defined that A κ-contains B, or $A \succcurlyeq B$. *A properly* contains B ($A \succ B$) if the inequality is strict. Note that $ci(A, B) > ci(B, A)$ also implies that $|A| > |B|$, so circular relations such as $A \succ B \succ C \succ A$ are impossible; therefore, the corresponding graph representation (see below) becomes acyclic.

This generalized definition of containment coincides with the conventional one when $\kappa = 1$, that is $A \subseteq B$ if every element of B belongs to A. However, when $\kappa < 1$, the requirement for defining containment is relaxed. As a result, we may still declare that A contains B even if some

elements of B are not part of A, to a degree. This is to account for the ambiguous classification of data points at the decision boundaries of the clustering algorithm.

In practice, A and B are represented by binary ownership vectors. *Pairwise containment indices* of communities are calculated via the following formula:

$$\tilde{\mathbf{C}} = \mathbf{P}^T\mathbf{P},$$

$$\mathbf{C} = \left[diag\left(\tilde{\mathbf{C}}\right)\right]^{-1}\tilde{\mathbf{C}}, \tag{B. 3}$$

where $diag(\tilde{\mathbf{C}})$ is a diagonal matrix composed of diagonal elements of $\tilde{\mathbf{C}}$. Finally, $\mathbf{C}$ is the matrix of pairwise containment indices where each element $\mathbf{C}_{ij} = ci(\boldsymbol{p}_i, \boldsymbol{p}_j)$.

Containment graph. Let G be a directed graph. V are the vertices, each corresponds to a cluster of data points. E are the directed edges. Let u and v be two vertices in G, and U and V their corresponding clusters. There exists an edge $u \to v$ if U κ-contains V. Data points themselves are treated as clusters of size 1 and represented by *terminal vertices*, which are connected to other vertices in the same way as the *non-terminal* vertices. All vertices receive an edge from the *root vertex* representing a grand cluster that owns every data point. The containment graph is acyclic as no circular relations are allowed (see above).

Since G represents all the *containment relations* among the clusters, it is referred to as a *containment graph*. Most of the containment relations are transitive, meaning that if there is an edge $v \to u$ and $u \to w$, there may be also an edge $v \to w$. The transitive reduction of G, denoted by H, is a subgraph of G that removes as many edges as possible while retaining the same reachability for every vertex, which effectively eliminates transitive edges, e.g. $v \to w$ in the above example. Since G is acyclic, H is unique.

Computationally, the *affinity matrix* of G can be easily evaluated from the matrix of pairwise containment indices, $\mathbf{C}$ (see above), by comparing it with κ:

$$\mathbf{A}(G) = \mathbf{C} \circ \Delta(\mathbf{C} \geq \kappa) \circ \Delta(\mathbf{C} \geq \mathbf{C}^T) \tag{B.4}$$

Here $\circ$ denotes the Hadamard (element-wise) product and Δ is the vectorized Kronecker delta function that operates on the input matrix element-wise. If there is any two-way edges, one of the edges is removed arbitrarily. V are sorted in a topological order $\alpha(V)$ and a reachability matrix, $\mathbf{R}(G) = [R_{ij}]$ where $i, j \in \alpha(V)$, is computed for the convenience of the latter operations. $\alpha(V)$ and $R(G)$ can be computed together in a single run of depth-first search (DFS). The transitive reduction H of G was found with the following algorithm,

Algorithm 1 Transitive Reduction

Input: a containment graph $G = (V, E)$
Output: a directed acyclic graph H
foreach $u \in V$ **do**
| **foreach** $v \in direct\_descendant(u)$ **do**
| | **foreach** $w \in descendant(v)$ **do**
| | | remove (u, w) if $(u, w) \in E$
| | **end**
| **end**
end

$direct\_descendant(u)$ finds all the vertices that receive an edge from u ; whereas $descendant(v)$ finds all the vertices *reachable* from v, which can be efficiently found using the

adjacency or the reachability matrix, respectively. Notably, the topological order $\alpha(V)$ and the reachability matrix $\mathbf{R}(G)$ stays valid for H.

Every vertex in H except for the root can have one to several direct predecessors. Consider a vertex and all the edges incident to it. The edge with the highest weight is considered as an *essential edge* for maintaining reachability of the vertex; whereas the others are considered as *competing edges*. Competing edges are ranked globally by their weights, such that a second cutoff value $\xi \in [0, 1]$ can be used to select top ranking ones. This produces a subgraph of H, denoted by $\widetilde{H}$, containing all the essential edges and the top $\xi \times 100\%$ competing edges. Notably, there are two special cases: when $\xi = 1$, $\widetilde{H} = H$; when $\xi = 0$, $\widetilde{H}$ is an *arborescence*, which means there exists only a single (directed) path from the root to any vertex in the graph. In this study, $\xi = 0$ was used.

Finally, to further simplify $\widetilde{H}$, vertices with both indegree and outdegree equal to 1 are removed by contracting the edge between them and their neighbors. While the topological order of the remaining vertices $\alpha(\tilde{V})$ is consistent with $\alpha(V)$, $\mathbf{R}(G) = \mathbf{R}(H) \neq \mathbf{R}(\widetilde{H})$.

Vertex (cluster) depth. The *depth* of a vertex is defined as the max number of steps for it to traverse back to the root. It is effectively computed by the following algorithm:

Algorithm 2 Computing Depths

Input: simplified containment graph $\widetilde{H}$ and the topological order of vertices $\alpha(\tilde{V})$
Output: an array $\boldsymbol{d}$ containing the depths of $\alpha(\tilde{V})$
initialize $\boldsymbol{d}$ as an integer array of length
foreach $u \in \alpha(V)$
    if u is root **then**
        $\boldsymbol{d}_u = 0$

```
    else
        p = get_upper_depths(u)
        d_u = max(p) + 1
    end
end
```

$get\_upper\_depths(u)$ returns the depths of all the direct predecessors of vertex u. This algorithm works based on the assumption that when a vertex is being evaluated, all of its predecessors must have been evaluated already, which is conveniently satisfied by the topological sorting. The algorithm tends to assign the largest possible depth to the vertices; however, this behavior can be easily modified to fulfill different purposes (for example, one may be interested in the "shallowness" of the clusters).

Now that we have a containment graph $\widetilde{H}$ where each vertex belongs to an original cluster, the ownership of the clusters, i.e. the vertices that belong to that cluster, may be subject to change. *Adjusted ownership* for a non-terminal vertex is defined as all its reachable terminal vertices. The smaller the values of κ and/or ξ, the larger the difference between the original and the new clusters. Let ω_i be a terminal vertex representing the data point x_i, and $\mu_1, \dots \mu_{\widetilde{m}}$ be $\widetilde{m}$ non-terminal vertices at a given depth. An $n \times \widetilde{m}$ binary matrix, $\mathbf{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_{\widetilde{m}}] = [q_{ij}]$, can be used to describe the membership of adjusted clusters,

$$q_{ij} = \begin{cases} 1 & \text{if } \omega_i \text{ is reachable from } \mu_j, \\ 0 & \text{otherwise.} \end{cases} \tag{B.5}$$

Bibliography

Alpert, C.J., Kahng, A.B. and Yao, S.-Z. Spectral partitioning with multiple eigenvectors. *Discrete Appl Math* 1999;90(1-3):3-26.

Alpert, C.J. and Yao, S.-Z. Spectral partitioning: the more eigenvectors, the better. In, *Proceedings of the 32nd annual ACM/IEEE Design Automation Conference*. 1995. p. 195-200.

Amadei, A., Linssen, A.B. and Berendsen, H.J. Essential dynamics of proteins. *Proteins* 1993;17(4):412-425.

An, L., *et al.* OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries. *Genome Biol* 2019;20(1):282.

Andrey, G. and Mundlos, S. The three-dimensional genome: regulating gene expression during pluripotency and development. *Development* 2017;144(20):3646-3658.

Atilgan, A.R., *et al.* Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal* 2001;80(1):505-515.

Atilgan, C. and Atilgan, A.R. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput Biol* 2009;5(10):e1000544.

Ay, F., Bailey, T.L. and Noble, W.S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res* 2014;24(6):999-1011.

Ay, F., *et al.* Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res* 2014;24(6):974-988.

Babu, M.M. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem Soc Trans* 2016;44(5):1185-1200.

Bahar, I., *et al.* Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability. *Phys Rev Lett* 1998;80(12):2733.

Bahar, I., Atilgan, A.R. and Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design* 1997;2(3):173-181.

Bahar, I., *et al.* Structure-Encoded Global Motions and Their Role in Mediating Protein-Substrate Interactions. *Biophys J* 2015;109(6):1101-1109.

Bahar, I., Chennubhotla, C. and Tobi, D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr Opin Struct Biol* 2007;17(6):633-640.

Bahar, I., Jernigan, R.L. and Dill, K.A. Protein Actions: Principles and Modeling by I. Bahar, RL Jernigan, and KA Dill. In.: Springer; 2017.

Bahar, I., *et al.* Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem Rev* 2010;110(3):1463-1497.

Bahar, I., *et al.* Global dynamics of proteins: bridging between structure and function. *Annu Rev Biophys* 2010;39:23-42.

Bahar, I. and Rader, A.J. Coarse-grained normal mode analysis in structural biology. *Curr Opin Struct Biol* 2005;15(5):586-592.

Bahar, I., *et al.* Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. *Biochemistry* 1998;37(4):1067-1075.

Bakan, A. and Bahar, I. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc Natl Acad Sci U S A* 2009;106(34):14349-14354.

Bakan, A., *et al.* Evol and ProDy for bridging protein sequence evolution and structural dynamics. *Bioinformatics* 2014;30(18):2681-2683.

Bakan, A., Meireles, L.M. and Bahar, I. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* 2011;27(11):1575-1577.

Banner, D.W., *et al.* Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 angstrom resolution using amino acid sequence data. *Nature* 1975;255(5510):609-614.

Barrett, T., *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* 2013;41(Database issue):D991-995.

Batista, P.R., *et al.* Free Energy Profiles along Consensus Normal Modes Provide Insight into HIV-1 Protease Flap Opening. *J Chem Theory Comput* 2011;7(8):2348-2352.

Batista, P.R., *et al.* Consensus modes, a robust description of protein collective motions from multiple-minima normal mode analysis--application to the HIV-1 protease. *Phys Chem Chem Phys* 2010;12(12):2850-2859.

Bau, D. and Marti-Renom, M.A. Structure determination of genomic domains by satisfaction of spatial restraints. *Chromosome Res* 2011;19(1):25-35.

Bau, D., *et al.* The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* 2011;18(1):107-114.

Berman, H.M., *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000;28(1):235-242.

Bernstein, B.E., Meissner, A. and Lander, E.S. The mammalian epigenome. *Cell* 2007;128(4):669-681.

Bickmore, W.A. and van Steensel, B. Genome architecture: domain organization of interphase chromosomes. *Cell* 2013;152(6):1270-1284.

Blondel, V.D., *et al.* Fast unfolding of communities in large networks. *J Stat Mech: Theory Exp* 2008;2008(10):P10008.

Bonev, B., *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* 2017;171(3):557-572 e524.

Brüschweiler, R. Collective protein dynamics and nuclear spin relaxation. *J Chem Phys* 1995;102(8):3396-3403.

Buenrostro, J.D., *et al.* Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 2013;10(12):1213-1218.

Carnevale, V., *et al.* Convergent dynamics in the protease enzymatic superfamily. *J Am Chem Soc* 2006;128(30):9766-9772.

Cattoglio, C., *et al.* Determining cellular CTCF and cohesin abundances to constrain 3D genome models. *Elife* 2019;8.

Cavalli, G. and Misteli, T. Functional implications of genome topology. *Nat Struct Mol Biol* 2013;20(3):290-299.

Chan, P.K., Schlag, M.D. and Zien, J.Y. Spectral k-way ratio-cut partitioning and clustering. *IEEE Trans Comput-Aided Des Integr Circuits Syst* 1994;13(9):1088-1096.

Chen, E.Y., *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013;14:128.

Chen, H., *et al.* Functional organization of the human 4D Nucleome. *Proc Natl Acad Sci U S A* 2015;112(26):8002-8007.

Chen, J., Hero, A.O., 3rd and Rajapakse, I. Spectral identification of topological domains. *Bioinformatics* 2016;32(14):2151-2158.

Cheng, M.H. and Bahar, I. Complete mapping of substrate translocation highlights the role of LeuT N-terminal segment in regulating transport cycle. *PLoS Comput Biol* 2014;10(10):e1003879.

Cheng, M.H. and Bahar, I. Molecular Mechanism of Dopamine Transport by Human Dopamine Transporter. *Structure* 2015;23(11):2171-2181.

Consortium, E.P. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;306(5696):636-640.

Cremer, T. and Cremer, M. Chromosome territories. *Cold Spring Harb Perspect Biol* 2010;2(3):a003889.

Cui, Q. and Bahar, I. Normal mode analysis: theory and applications to biological and chemical systems. CRC press; 2005.

Darrow, E.M., *et al.* Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc Natl Acad Sci U S A* 2016;113(31):E4504-4512.

Dawson, N.L., *et al.* CATH-Gene3D: Generation of the Resource and Its Use in Obtaining Structural and Functional Annotations for Protein Sequences. *Methods Mol Biol* 2017;1558:79-110.

Dixon, J.R., *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* 2015;518(7539):331-336.

Dixon, J.R., *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;485(7398):376-380.

Doruker, P., Atilgan, A.R. and Bahar, I. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α - amylase inhibitor. *Proteins* 2000;40(3):512-524.

Dowen, J.M., *et al.* Multiple structural maintenance of chromosome complexes at transcriptional regulatory elements. *Stem Cell Rep* 2013;1(5):371-378.

Drew, D. and Boudker, O. Shared Molecular Mechanisms of Membrane Transporters. *Annu Rev Biochem* 2016;85:543-572.

Duan, Z., *et al.* A three-dimensional model of the yeast genome. *Nature* 2010;465(7296):363-367.

Durand, N.C., *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* 2016;3(1):95-98.

Echave, J., Spielman, S.J. and Wilke, C.O. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet* 2016;17(2):109-121.

Echave, J. and Wilke, C.O. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. *Annu Rev Biophys* 2017;46:85-103.

El-Gebali, S., *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* 2019;47(D1):D427-D432.

Eyal, E. and Bahar, I. Toward a molecular understanding of the anisotropic response of proteins to external forces: insights from elastic network models. *Biophys J* 2008;94(9):3424-3435.

Faham, S., *et al.* The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na+/sugar symport. *Science* 2008;321(5890):810-814.

Filippova, D., *et al.* Identification of alternative topological domains in chromatin. *Algorithms Mol Biol* 2014;9:14.

Forcato, M., *et al.* Comparison of computational methods for Hi-C data analysis. *Nat Methods* 2017;14(7):679-685.

Forman-Kay, J.D. and Mittag, T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure* 2013;21(9):1492-1499.

Fraser, J., *et al.* An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiol Mol Biol Rev* 2015;79(3):347-372.

Gao, X., *et al.* Structure and mechanism of an amino acid antiporter. *Science* 2009;324(5934):1565-1568.

Gao, X., *et al.* Mechanism of substrate recognition and transport by an amino acid antiporter. *Nature* 2010;463(7282):828-832.

Gilson, A.I., *et al.* The Role of Evolutionary Selection in the Dynamics of Protein Structure Evolution. *Biophys J* 2017;112(7):1350-1365.

Gomez-Diaz, E. and Corces, V.G. Architectural proteins: regulators of 3D genome organization in cell fate. *Trends Cell Biol* 2014;24(11):703-711.

Grossfield, A., Feller, S.E. and Pitman, M.C. Convergence of molecular dynamics simulations of membrane proteins. *Proteins* 2007;67(1):31-40.

Gur, M., *et al.* Energy landscape of LeuT from molecular simulations. *J Chem Phys* 2015;143(24):243134.

Hagen, L. and Kahng, A.B. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans Comput-Aided Des Integr Circuits Syst* 1992;11(9):1074-1085.

Haliloglu, T. and Bahar, I. Adaptability of protein structures to enable functional interactions and evolutionary implications. *Curr Opin Struct Biol* 2015;35:17-23.

Haliloglu, T., Bahar, I. and Erman, B. Gaussian dynamics of folded proteins. *Phys Rev Lett* 1997;79(16):3090.

Hay, S.B., *et al.* The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp Hematol* 2018;68:51-61.

Heidari, N., *et al.* Genome-wide map of regulatory interactions in the human genome. *Genome Res* 2014;24(12):1905-1917.

Hess, B. Convergence of sampling in protein simulations. *Phys Rev E Stat Nonlin Soft Matter Phys* 2002;65(3 Pt 1):031910.

Hinsen, K., *et al.* Harmonicity in slow protein dynamics. *Chemical Physics* 2000;261(1-2):25-37.

Hollup, S.M., *et al.* Exploring the factors determining the dynamics of different protein folds. *Protein Sci* 2011;20(1):197-209.

Holm, L. and Laakso, L.M. Dali server update. *Nucleic Acids Res* 2016;44(W1):W351-355.

Holm, L. and Rosenstrom, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res* 2010;38(Web Server issue):W545-549.

Hou, C., *et al.* Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell* 2012;48(3):471-484.

Ilyin, V.A., *et al.* ModView, visualization of multiple protein sequences and structures. *Bioinformatics* 2003;19(1):165-166.

Jin, F., *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 2013;503(7475):290-294.

Jin, R., *et al.* Crystal structure and association behaviour of the GluR2 amino-terminal domain. *EMBO J* 2009;28(12):1812-1823.

Joeng, M., *et al.* A cell type-specific class of chromatin loops anchored at large DNA methylation nadirs. *bioRxiv* 2017:212928.

Juritz, E., *et al.* Protein conformational diversity modulates sequence divergence. *Mol Biol Evol* 2013;30(1):79-87.

Kalayil, S., Schulze, S. and Kuhlbrandt, W. Arginine oscillation explains Na+ independence in the substrate/product antiporter CaiT. *Proc Natl Acad Sci U S A* 2013;110(43):17296-17301.

Kaynak, B.T., Findik, D. and Doruker, P. RESPEC Incorporates Residue Specificity and the Ligand Effect into the Elastic Network Model. *J Phys Chem B* 2018;122(21):5347-5355.

Kazmier, K., Claxton, D.P. and McHaourab, H.S. Alternating access mechanisms of LeuT-fold transporters: trailblazing towards the promised energy landscapes. *Curr Opin Struct Biol* 2017;45:100-108.

Kazmier, K., *et al.* Conformational cycle and ion-coupling mechanism of the Na+/hydantoin transporter Mhp1. *Proc Natl Acad Sci U S A* 2014;111(41):14752-14757.

Kent, W.J., *et al.* The human genome browser at UCSC. *Genome Res* 2002;12(6):996-1006.

Klemm, S.L., Shipony, Z. and Greenleaf, W.J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 2019;20(4):207-220.

Knight, P.A. and Ruiz, D. A fast algorithm for matrix balancing. *IMA J Numer Anal* 2013;33(3):1029-1047.

Knudsen, M. and Wiuf, C. The CATH database. *Hum Genomics* 2010;4(3):207-212.

Kodama, Y., *et al.* The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* 2012;40(Database issue):D54-56.

Koshy, C. and Ziegler, C. Structural insights into functional lipid-protein interactions in secondary transporters. *Biochim Biophys Acta* 2015;1850(3):476-487.

Krieger, J., Bahar, I. and Greger, I.H. Structure, Dynamics, and Allosteric Potential of Ionotropic Glutamate Receptor N-Terminal Domains. *Biophys J* 2015;109(6):1136-1148.

Krishnamurthy, H. and Gouaux, E. X-ray structures of LeuT in substrate-free outward-open and apo inward-open states. *Nature* 2012;481(7382):469-474.

Kuhn, H.W. The Hungarian method for the assignment problem. *Nav Res Logist Q* 1955;2(1‐2):83-97.

Kuhn, H.W. Variants of the Hungarian method for assignment problems. *Nav Res Logist Q* 1956;3(4):253-258.

Kuleshov, M.V., *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;44(W1):W90-97.

Kurkcuoglu, Z., Bahar, I. and Doruker, P. ClustENM: ENM-Based Sampling of Essential Conformational Space at Full Atomic Resolution. *J Chem Theory Comput* 2016;12(9):4549-4562.

Lachmann, A., *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun* 2018;9(1):1366.

Lajoie, B.R., Dekker, J. and Kaplan, N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 2015;72:65-75.

Letunic, I. and Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 2007;23(1):127-128.

Letunic, I. and Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47(W1):W256-W259.

Levy-Leduc, C., *et al.* Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* 2014;30(17):i386-392.

Lezon, T.R. and Bahar, I. Constraints imposed by the membrane selectively guide the alternating access dynamics of the glutamate transporter GltPh. *Biophys J* 2012;102(6):1331-1340.

Li, T., *et al.* A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat Methods* 2017;14(1):61-64.

Liberles, D.A., *et al.* The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci* 2012;21(6):769-785.

Lieberman-Aiden, E., *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;326(5950):289-293.

Liu, Y. and Bahar, I. Sequence evolution correlates with structural dynamics. *Mol Biol Evol* 2012;29(9):2253-2263.

Maguid, S., Fernandez-Alberti, S. and Echave, J. Evolutionary conservation of protein vibrational dynamics. *Gene* 2008;422(1-2):7-13.

Maguid, S., *et al.* Evolutionary conservation of protein backbone flexibility. *J Mol Evol* 2006;63(4):448-457.

Malinauskaite, L., *et al.* A mechanism for intracellular release of Na+ by neurotransmitter/sodium symporters. *Nat Struct Mol Biol* 2014;21(11):1006-1012.

Meaburn, K.J. and Misteli, T. Cell biology: chromosome territories. *Nature* 2007;445(7126):379-781.

Meilă, M. Comparing clusterings by the variation of information. In, *Learning theory and kernel machines*. Springer; 2003. p. 173-187.

Merkle, P.S., *et al.* Substrate-modulated unwinding of transmembrane helices in the NSS transporter LeuT. *Sci Adv* 2018;4(5):eaar6179.

Mikulska-Ruminska, K., *et al.* Characterization of Differential Dynamics, Specificity, and Allostery of Lipoxygenase Family Members. *J Chem Inf Model* 2019;59(5):2496-2508.

Mirny, L.A. The fractal globule as a model of chromatin architecture in the cell. *Chromosome research* 2011;19(1):37-51.

Nagano, T., *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;502(7469):59-64.

Neph, S., *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* 2012;28(14):1919-1920.

Nevin Gerek, Z., Kumar, S. and Banu Ozkan, S. Structural dynamics flexibility informs function and evolution at a proteome scale. *Evol Appl* 2013;6(3):423-433.

Oluwadare, O., Highsmith, M. and Cheng, J. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biol Proced Online* 2019;21:7.

Patro, R., *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;14(4):417-419.

Pellin, D., *et al.* A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat Commun* 2019;10(1):2395.

Penmatsa, A., Wang, K.H. and Gouaux, E. X-ray structure of dopamine transporter elucidates antidepressant mechanism. *Nature* 2013;503(7474):85-90.

Perez, C., *et al.* Alternating-access mechanism in conformationally asymmetric trimers of the betaine transporter BetP. *Nature* 2012;490(7418):126-130.

Perica, T., *et al.* Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science* 2014;346(6216):1254346.

Phanstiel, D.H., *et al.* Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Mol Cell* 2017;67(6):1037-1048 e1036.

Ponzoni, L., *et al.* Shared dynamics of LeuT superfamily members and allosteric differentiation by structural irregularities and multimerization. *Philos Trans R Soc Lond B Biol Sci* 2018;373(1749).

Porto, M., *et al.* Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. *Mol Biol Evol* 2005;22(3):630-638.

Prim, R.C. Shortest connection networks and some generalizations. *Bell system technical journal* 1957;36(6):1389-1401.

Rao, S.S., *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;159(7):1665-1680.

Rao, S.S.P., *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* 2017;171(2):305-320 e324.

Redfern, O.C., Dessailly, B. and Orengo, C.A. Exploring the structure and function paradigm. *Curr Opin Struct Biol* 2008;18(3):394-402.

Reichardt, J. and Bornholdt, S. Statistical mechanics of community detection. *Phys Rev E Stat Nonlin Soft Matter Phys* 2006;74(1 Pt 2):016110.

Ressl, S., *et al.* Molecular basis of transport and regulation in the Na(+)/betaine symporter BetP. *Nature* 2009;458(7234):47-52.

Romo, T.D. and Grossfield, A. Block Covariance Overlap Method and Convergence in Molecular Dynamics Simulation. *J Chem Theory Comput* 2011;7(8):2464-2472.

Rosvall, M., Axelsson, D. and Bergstrom, C.T. The map equation. *Eur Phys J* 2009;178(1):13-23.

Rousseau, M., *et al.* Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics* 2011;12:414.

Rowley, M.J. and Corces, V.G. Organizational principles of 3D genome architecture. *Nat Rev Genet* 2018;19(12):789-800.

Rudan, M.V., *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell reports* 2015;10(8):1297-1309.

Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4(4):406-425.

Sanborn, A.L., *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* 2015;112(47):E6456-6465.

Sanyal, A., *et al.* The long-range interaction landscape of gene promoters. *Nature* 2012;489(7414):109-113.

Sauerwald, N. and Kingsford, C. Quantifying the similarity of topological domains across normal and cancer human cell types. *Bioinformatics* 2018;34(13):i475-i483.

Sauerwald, N., *et al.* Chromosomal dynamics predicted by an elastic network model explains genome-wide accessibility and long-range couplings. *Nucleic Acids Res* 2017;45(7):3663-3673.

Schulze, S., *et al.* Structural basis of Na(+)-independent and cooperative substrate/product antiport in CaiT. *Nature* 2010;467(7312):233-236.

Serra, F., *et al.* Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol* 2017;13(7):e1005665.

Shi, Y. Common folds and transport mechanisms of secondary active transporters. *Annu Rev Biophys* 2013;42:51-72.

Shimamura, T., *et al.* Molecular basis of alternating access membrane transport by the sodium-hydantoin transporter Mhp1. *Science* 2010;328(5977):470-473.

Shindyalov, I.N. and Bourne, P.E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 1998;11(9):739-747.

Singh, S.K., *et al.* A competitive inhibitor traps LeuT in an open-to-out conformation. *Science* 2008;322(5908):1655-1661.

Singh, S.K., Yamashita, A. and Gouaux, E. Antidepressant binding site in a bacterial homologue of neurotransmitter transporters. *Nature* 2007;448(7156):952-956.

Skjaerven, L., Reuter, N. and Martinez, A. Dynamics, flexibility and ligand-induced conformational changes in biological macromolecules: a computational approach. *Future Med Chem* 2011;3(16):2079-2100.

Sokal, R.R. A statistical method for evaluating systematic relationships. *Univ Kansas, Sci Bull* 1958;38:1409-1438.

Speicher, M.R. and Carter, N.P. The new cytogenetics: blurring the boundaries with molecular biology. *Nat Rev Genet* 2005;6(10):782-792.

Stella, X.Y. and Shi, J. Multiclass spectral clustering. In, *Proc IEEE Int Conf Comput Vis*. IEEE; 2003. p. 313.

Stevens, T.J., *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 2017;544(7648):59-64.

Stuart, T., *et al.* Comprehensive Integration of Single-Cell Data. *Cell* 2019;177(7):1888-1902 e1821.

Szklarczyk, D., *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;47(D1):D607-D613.

Theobald, D.L. and Wuttke, D.S. Divergent evolution within protein superfolds inferred from profile-based phylogenetics. *J Mol Biol* 2005;354(3):722-737.

Tirion, M.M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* 1996;77(9):1905.

Tokuriki, N. and Tawfik, D.S. Protein dynamism and evolvability. *Science* 2009;324(5924):203-207.

Tsai, C.J., *et al.* Folding funnels, binding funnels, and protein function. *Protein Sci* 1999;8(6):1181-1190.

Tsompana, M. and Buck, M.J. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 2014;7(1):33.

UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;47(D1):D506-D515.

Van Bortle, K., *et al.* Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol* 2014;15(6):R82.

Varoquaux, N., *et al.* A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* 2014;30(12):i26-33.

Wang, K.H., Penmatsa, A. and Gouaux, E. Neurotransmitter and psychostimulant recognition by the dopamine transporter. *Nature* 2015;521(7552):322-327.

Watanabe, A., *et al.* The mechanism of sodium and substrate release from the binding pocket of vSGLT. *Nature* 2010;468(7326):988-991.

Weinreb, C. and Raphael, B.J. Identification of hierarchical chromatin domains. *Bioinformatics* 2016;32(11):1601-1609.

Weyand, S., *et al.* Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter. *Science* 2008;322(5902):709-713.

Xu, Y., *et al.* Induced-fit or preexisting equilibrium dynamics? Lessons from protein crystallography and MD simulations on acetylcholinesterase and implications for structure-based drug design. *Protein Sci* 2008;17(4):601-605.

Xu, Z., *et al.* FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. *Bioinformatics* 2016;32(17):2692-2695.

Yaffe, E. and Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* 2011;43(11):1059-1065.

Yamashita, A., *et al.* Crystal structure of a bacterial homologue of Na+/Cl--dependent neurotransmitter transporters. *Nature* 2005;437(7056):215-223.

Yan, K.K., Lou, S. and Gerstein, M. MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. *PLoS Comput Biol* 2017;13(7):e1005647.

Yates, A.D., *et al.* Ensembl 2020. *Nucleic Acids Res* 2020;48(D1):D682-D688.

Yuen, K.C., Slaughter, B.D. and Gerton, J.L. Condensin II is anchored by TFIIIC and H3K4me3 in the mammalian genome and supports the expression of active dense gene clusters. *Sci Adv* 2017;3(6):e1700191.

Zhan, Y., *et al.* Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res* 2017;27(3):479-490.

Zhang, B. and Wolynes, P.G. Topology, structures, and energy landscapes of human chromosomes. *Proc Natl Acad Sci U S A* 2015;112(19):6062-6067.

Zhang, S., Chen, F. and Bahar, I. Differences in the intrinsic spatial dynamics of the chromatin contribute to cell differentiation. *Nucleic Acids Res* 2020;48(3):1131-1145.

Zhang, S., *et al.* Shared Signature Dynamics Tempered by Local Fluctuations Enables Fold Adaptability and Specificity. *Mol Biol Evol* 2019;36(9):2053-2068.

Zhang, Y., *et al.* Intrinsic dynamics is evolutionarily optimized to enable allosteric behavior. *Curr Opin Struct Biol* 2020;62:14-21.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;57(4):702-710.

Zheng, W. and Brooks, B.R. Probing the local dynamics of nucleotide-binding pocket coupled to the global dynamics: myosin versus kinesin. *Biophys J* 2005;89(1):167-178.

Zheng, W., Brooks, B.R. and Thirumalai, D. Allosteric transitions in biological nanomachines are described by robust normal modes of elastic networks. *Curr Protein Pept Sci* 2009;10(2):128-132.

Zomot, E., Gur, M. and Bahar, I. Microseconds simulations reveal a new sodium-binding site and the mechanism of sodium-coupled substrate uptake by LeuT. *J Biol Chem* 2015;290(1):544-555.