Comparing the Intrinsic Dynamics of Multiple Protein Structures Using Elastic Network Models

Edvin Fuglebakk[1,2], Sandhya P. Tiwari [1,2], Nathalie Reuter [1,2*]

[1] Department of Molecular Biology, University of Bergen, Pb. 7803, N-5020 Bergen, Norway

[2] Computational Biology Unit, Department of Informatics, University of Bergen, Pb. 7803, N-5020 Bergen, Norway

E.F: Edvin.Fuglebakk@mbi.uib.no

S.T.: Sandhya.Tiwari@mbi.uib.no

* To whom correspondence should be addressed:

Nathalie Reuter, University of Bergen, Department of Molecular Biology, Pb. 7803, N-5020 Bergen, Norway

Tel: (+47) 555 84040 // Fax: (+47) 555 89683

E-mail address: nathalie.reuter@mbi.uib.no

ABSTRACT (max 250 words, need to contain the following categories)

Background. Elastic Network Models (ENMs) are based on the simple idea that a protein can be described as a set of particles connected by springs, which can then be used to describe its intrinsic flexibility using, for example, normal mode analysis. Since the introduction of the first ENM by Monique Tirion in 1996, several variants using coarser protein models have been proposed and their reliability for the description of proteins intrinsic dynamics has been widely demonstrated. Lately an increasing number of studies have focused on the meaning of slow dynamics for proteins function and its potential conservation through evolution, naturally leading to comparisons of the intrinsic dynamics of multiple protein structures with varying levels of similarity.

Scope of Review. We describe computational strategies for calculating and comparing intrinsic dynamics of multiple proteins using elastic network models, as well as a selection of examples from the recent literature.

Major Conclusions. The increasing interest for comparing dynamics across proteins structures with various levels of similarity, has led to the establishment and validation of reliable computational strategies using ENMs. Comparing dynamics has been shown to be a viable way for gaining greater understanding for the mechanisms employed by proteins for their function. Choices of ENM parameters, structure alignment or similarity measures will likely influence the interpretation of dynamics comparative analysis.

General Significance. Understanding the relation between protein function and dynamics is relevant to the fundamental understanding of proteins structure-dynamics-function relationship.

KEYWORDS: elastic network models, protein dynamics, intrinsic dynamics, normal mode analysis

ABBREVIATIONS

ENM: elastic network model

NMA

PCA

NMR

RMSIP

RMSD

1. Introduction

Folded proteins are remarkably dense, with a heterogeneously distributed density, which reflects the uneven distribution of interatomic forces in the protein. Their response to thermal forces is expected to proceed by preferably deforming the least compact regions while keeping the most compact ones rigid. Atoms tightly coupled on short time-scales are expected to remain tightly coupled on longer time-scales, at least between unfolding events. This suggests that estimates of the atomic density distribution within a folded protein can capture its collective degrees of freedom. It also motivates the extrapolation from analysis of intrinsic properties of the structure to collective motions occurring on for example the millisecond time-scale. Estimates of the atomic density distribution can also replace information about the exact chemistry involved in stabilizing the fold, similar to how the elastic response of macroscopic materials can be calculated without atomic detail.

Likewise Elastic Network Models (ENMs) are based on the simple idea that a protein can be described as a set of particles connected by springs, which can then be used to describe its intrinsic flexibility using normal mode analysis (NMA). Monique Tirion pioneered the field in 1996, when she showed that a single-parameter potential could reproduce the slow dynamics obtained with a more complicated potential {Tirion, 1996 #42}. This simplification makes the potential insensitive to the details of the equilibrium structure, which has minimal energy by construction. Models from experimental structure determination can thus be used directly, without the costly 4energy minimization associated with the use of chemical force fields. Tirion's model has later been further simplified, in particular increasing its coarseness, such as in ENMs of interacting residues, rather than atoms {Hinsen, 1998 #2;Atilgan, 2001 #3}. ENMs provide a simple and interpretable description of the proteins collective motion, can be conveniently coarse grained, and are computationally inexpensive to calculate. For these reasons they rapidly have replaced molecular mechanics force fields that had been, since as early as 1977, used for NMA of proteins {Brooks, 1985 #7;Go, 1983 #8;Levitt, 1985 #9;McCammon, 1977 #6;Noguti, 1982 #10;Brooks, 1983 #5}.

The robustness of NMA with ENMs for the description of protein's slow collective motions has almost come as a surprise to some. The motivation outlined above for using ENMs involved some brave assumptions, and it was not *a priori* obvious that

these assumptions were valid: the harmonic approximation used for investigating dynamics of large conformational changes and the absence of frictions such as those caused by the solvent. Yet early studies comparing NMA and experimental structural data, or molecular dynamics simulations did validate the use of NMA with coarse-grained model. Validation against detailed molecular mechanics force fields on large protein datasets have shown that even coarser models than the one suggested by Tirion still reproduce the slow dynamics obtained from molecular simulations (e.g.{Pontiggia, 2007 #42;Rueda, 2007 #40;Skjaerven, 2010 #43;Yang, 2008 #41}). Furthermore, several studies have shown that in many cases a few low-energy normal modes account for most of the structure difference between two conformational states {Marques, 1995 #960;Hinsen, 1999 #138;Tama, 2001 #1040;Krebs, 2002 #926}. Conformational changes can be described by just a few low-energy normal modes intimately linked to the structure indicating that proteins systematically make use of these low energy modes to achieve their function. The importance of these modes for proteins function has naturally led to the question of the evolutionary conservation of proteins slow dynamics, in analogy to structure and sequence.

Examples of comparative dynamics analysis include studying a set of proteins that represent various functional states of a given enzyme upon e.g. ligand-binding {Seckler, 2013 #91; Rodgers, 2013 #57}, evaluating the conservation of dynamics within a homologous protein family{Kolan, 2014 #119;Laberge, 2007 #70;Lukman, 2009 #79;Maguid, 2005 #17;Marcos, 2011 #97;Raimondi, 2010 #23;Velazquez-Muriel, 2009 #22}, or within a set of proteins that possess the same fold despite low sequence identity{Zheng, 2003 #105;Hollup, 2011 #122}. In a recent article, Cristian Micheletti comprehensively reviews the use of dynamics as an aid for sequence and structure alignments of proteins {Micheletti, 2013 #44}. It has been shown, when comparing structures of homologous proteins and their intrinsic dynamics, that protein structures evolve along the low energy modes. A number of studies have shown that the low energy modes are robust to sequence variations {Nicolay, 2006 #16;Tama, 2006 #110;Echave, 2010 #25;Hollup, 2011 #122; Yang, 2008 #51;Zheng, 2007 #43;Zheng, 2006 #54}. The use of ENMs for comparative protein dynamics has the potential to teach us more on a wide range of topics. To name a few, these can be about the effects of ligand binding, whether in an active site or an allosteric site,

changes in oligomeric state, changes in sequence through evolution, and the level of similarity in dynamics between functionally similar enzymes.

Together with the question of the evolutionary conservation of internal dynamics has come the need to reliably compare computed dynamics for a set of protein structures. Due to the scarcity of experimental data on protein dynamics, molecular modelling at large is an attractive alternative that has earlier demonstrated its predictive power through numerous applications. ENMs are a model of choice for such studies, even if computing power has admittedly become more affordable than it was at the advent of ENMs and molecular dynamics simulations on microseconds time-scale are becoming increasingly accessible to the research community. The tractability and simplicity of ENMs is unparalleled by molecular mechanics force fields and ENMs defined with transferrable parameters can be easily applied to large numbers of protein structures in automated ways. Beyond the choice of the ENM and its parameterization, comparing internal dynamics of a set of several protein structures comes with a set of methodological choices, which are not obvious, but can significantly affect the outcome of the comparative dynamics analysis. After an introduction to the formalism of ENMs and their parameterization, we focus on aspects that are directly relevant for comparative analysis of multiple protein structures such as the similarity measures used to compare computed dynamics, the influence of the alignment methods and how to include the influence of those regions of the structures that are not similar in sequence or conserved into the comparison. Next, using selected examples, we describe how comparing proteins intrinsic dynamics can be successfully used to understand conformational changes upon ligand binding, functional oligomerisation states and the role of intrinsic dynamics for proteins function. Finally we list some of the most commonly used software and libraries for ENM calculations.

2. Elastic Network Models

2.1. Formalism

Since Tirion's contribution {Tirion, 1996 #1}, further simplifications of the ENMs have been made. Upon realising that a good density estimate can be made even without atomic detail and that backbone motion can be largely decoupled from side-chain movement, Hinsen et al. {Hinsen, 1998 #2} introduced a model with non-uniform distance dependent force-constants, connecting only Cα atoms. Atilgan et al. {Atilgan, 2001 #3} also applied Tirion's model at the Cα granularity. Another popular density based model has been the early Gaussian Network Model (GNM) {Bahar, 1997 #4}. While obtaining density estimates in a way similar to Atilgan et al., this model does not employ a Hookean potential. The interpretation of GNMs is therefore different from the ENMs.

Since the initial ENMs, many variants have been proposed. More detailed descriptions of the local backbone configurations have been investigated, such as parameters dependent on the secondary structure of the backbone {Moritsugu, 2007 #5;Moritsugu, 2009 #6}, reintroduction of chemical bond information {Jeong, 2006 #112}{Kim, 2013 #52} as well as modelling of side-chain locations {Micheletti, 2004 #7}. On the other hand, simplification to fewer coordinates has been proposed, both in terms of simpler coordinate systems {Mendez, 2010 #8}{Wako, 2011 #50} and less granular representation of the proteins {Tama, 2000 #51}. Despite all this variety the ENMs can be understood in terms of a single unifying formalism, which will be detailed in the following.

The ENMs model the protein as a network of Hookean springs connecting all residues, which are typically represented by nodes located at the centre of their Cα atom. Interactions between atoms are described by the pair potential:

$$V_{ij}(\mathbf{r}) = \frac{k_{ij}}{2}\left(||\mathbf{r}_i - \mathbf{r}_j|| - ||\mathbf{r}_i^0 - \mathbf{r}_j^0||\right)^2 \qquad [1]$$

where $\mathbf{r}_i$ is the position of a residue i, in a configuration of the protein $\mathbf{r}$, the superscript 0 denotes the equilibrium conformation and k_{ij} is the force constant for the spring connecting residues i and j. Here k_{ij} is typically determined by a scalar function of distance between connected nodes. Apart from the choice of granularity of the

model, the function for determining k_{ij} is the most important difference between different ENMs. The potential energy of the entire network is the sum of this pair potential over all pairs:

$$V(\mathbf{r}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} V_{ij}(\mathbf{r}) \quad [2]$$

where N is the number of nodes in the network. Expanding this potential as a Taylor series around $\mathbf{r}^0$ reveals the following form of the potential:

$$V(\mathbf{r}) = \tfrac{1}{2} \left(\mathbf{r} - \mathbf{r}^0\right)^{\mathrm{T}} \mathbf{H} \left(\mathbf{r} - \mathbf{r}^0\right) \quad [3]$$

with **H** the matrix of partial second order derivatives of the potential. With respect to Cartesian coordinates this is a 3N × 3N matrix. The elements of **H** can be specified in terms of 3 × 3 submatrices corresponding to each pair of nodes:

$$\mathbf{H}_{ij} = \begin{cases} -\frac{k_{ij}}{||\mathbf{r}_i^0 - \mathbf{r}_j^0||^2} \left(\mathbf{r}_i^0 - \mathbf{r}_j^0\right) \left(\mathbf{r}_i^0 - \mathbf{r}_j^0\right)^{\mathrm{T}} & , i \neq j \\ \sum_{l \neq i} \mathbf{H}_{il} & , i = j \end{cases} \quad [4]$$

Since **H** is a symmetric matrix, the potential energy of a configuration ***r*** can be written in terms of its eigendecomposition:

$$V(\mathbf{r}) = \sum_{m=1}^{3N} \lambda_m \left(\left(\mathbf{r} - \mathbf{r}^0\right)^{\mathrm{T}} \mathbf{v}_m\right)^2 \quad [5]$$

Where $\mathbf{v}_m$ represents the normalized eigenvectors and λ_m the corresponding eigenvalues of **H**. These eigenvectors form an orthogonal basis for the configurational space of the protein so that they each provide energetically independent contributions to the potential energy of **r**. For small displacements from the energetic minimum, the displacement $\mathbf{r}\text{-}\mathbf{r}^0$ can be interpreted by via its decomposition into the eigenvectors $\mathbf{v}_m$ with high values for the inner product $(\mathbf{r}\text{-}\mathbf{r}^0)^{\mathrm{T}}\mathbf{v}_m$. These independent modes of deformation are referred to as the normal modes of the network, and they describe motion intrinsic to the protein structure. Motion along modes with high λ_m are intrinsically less energetically favourable than motion along lower energy modes. Because of the coarseness of the model eigenvalues are not interpreted exactly, but the separation between eigenvalues are informative of the relative energetic cost of different structural deformations. Since rigid-body rotations and translations of the network are not restrained, the six modes corresponding to rigid-body motion in

Cartesian coordinates will have zero energy. The modes describing rigid body displacements are referred to as trivial modes.

Since normal mode analysis has a long tradition in chemistry for analysing small vibrational molecules, the above formalism is often presented as an eigendecomposition of the mass weighted Hessian. In that case, the elastic network is considered a coupled harmonic oscillator and the eigenvalues are the squared frequencies of vibration along the corresponding modes. While the vibrational normal modes are a perfectly valid decomposition of motion, it is worth stressing that solvated proteins cannot in general be expected to be vibrational along their lower energy modes {Hinsen, 2008 #34} and thus, requires cautious interpretation of the oscillator model.

For equally normalized displacements, the quadratic dependence of energy on the spatial extent of deformations causes large local deformations to be more energetically expensive than collective motions that involve only small changes to each spring. Therefore low energy modes are expected to be collective. By a similar reasoning, collective motions can be expected to have larger amplitudes, as local deformations are constrained by the stronger local interactions. In fact, for a harmonic potential, the displacements along low-energy normal modes are exactly the deviations along high-variance principal components. The Boltzmann distribution for the potential given in Eq. [3] is a multivariate Gaussian distribution with a covariance matrix proportional to the inverse of **H**. Because of the zero energy associated with rigid movement of the protein, this inverse is not defined, but the Moore-Penrose pseudo-inverse **C**, can for many applications be regarded as a covariance matrix of internal deformations:

$$\mathbf{C} = \sum_{m=7}^{3N} \frac{1}{\lambda_m} \mathbf{v}_m \mathbf{v}_m^{\mathrm{T}} \qquad [6]$$

where the sum runs over the nontrivial modes. This implies that the eigenvectors $\mathbf{v}_m$ can be regarded as the principal components of this covariance matrix **C**, with variance $1/\lambda_m$. The covariance along each of the Cartesian coordinates of a pair of nodes *i* and *j* is proportional to C_{ij}, which denotes a 3 × 3 matrix. The trace of the submatrices C_{ii} is proportional to the mean squared thermal fluctuation of node *i*:

$$\langle ||\mathbf{r}_i - \mathbf{r}_i^0||^2 \rangle \propto \mathrm{tr}\,(\mathbf{C}_{ii}) = \sum_{m=7}^{3N} \frac{1}{\lambda_m} \mathbf{v}_m^{\mathrm{T}} \mathbf{v}_m \qquad [7]$$

where the angle brackets denote the mean and tr denotes the trace, or diagonal sum, of the matrix. To obtain a scalar quantification of the correlation of two nodes, a correlation matrix is commonly calculated, following Ichiye and Karplus {Ichiye, 1991 #9}:

$$P_{ij} = \frac{\mathrm{tr}(\mathbf{C}_{ij})}{(\mathrm{tr}(\mathbf{C}_{ii})\mathrm{tr}(\mathbf{C}_{jj}))^{\frac{1}{2}}} \quad [8]$$

Here the numerator is proportional to the expected inner product of displacement, which depends on both the magnitudes and the angles between node displacements, whenever $i \neq j$.

As mentioned above, the inner product in Eq. [5] quantifies the contribution of a mode to a small displacement from the energetic minimum. As a means to identify a few normal modes that approximate the displacement well, the squared overlap and related measures are commonly calculated {Marques, 1995 #686; Hinsen, 2000 #1255}. The squared overlap $\mathbf{O}_m$, of a normalized displacement vector $\mathbf{d}$ and a normal mode $\mathbf{v}_m$ is the squared inner product:

$$O_m(\mathbf{d}) = \left(\mathbf{d}^{\mathrm{T}}\mathbf{v}_m\right)^2 \quad [9]$$

with

$$\sum_{m=1}^{3N} O_m(\mathbf{d}) = 1 \quad [10]$$

since the normal modes are orthonormal. Such approaches are often applied even when the displacements analysed are not strictly infinitesimal. They have been important in validating the ENMs with experimentally determined displacements, as large overlaps with low energy modes indicates that a displacement is energetically favourable.

2.2. Parameterization: force constants and cut-offs

Apart from the choice of granularity and coordinate system used to represent the protein as an elastic network, the different ENMs proposed over the years mainly differ in how the force constants are determined (the function determining k_{ij} in Eq. [1]). While this function is commonly chosen to be a function of interatomic distance in the equilibrium conformation, model developers have not reached a consensus on

which mathematical formalism is more appropriate, or which benchmarking standards should be used. The simplest approach, following Tirion's initial model, uses uniform force constants for atoms or nodes that are within a given cutoff distance from each other. Springs longer than this cutoff is then assigned a force constant of zero, which is equivalent to just omitting the spring from the model. Other formalisms connect all nodes and set the force constants proportional to some function decaying with distance. Figure X illustrates the two approaches. Since energies of individual modes are typically only interpreted in normalized fashion, the exact values of the force constants are not important, only the contrast between them. While different choices of mathematical formalisms can be brought to close agreement through careful parameterization {Leioatts, 2012 #10}, it is important to choose appropriate benchmarks to parameterize against. Figure X illustrates the difference between common parameterisations of a uniform force constant model.

The parameterization of ENMs was initially motivated by comparison to detailed chemical potentials {Tirion, 1996 #1;Hinsen, 1998 #2} analysis of MD-trajectories {Hinsen, 2000 #35} and radial distribution analysis of the coordination between residues in the protein core {Atilgan, 2001 #3}. Taking advantage of the vast amount of structural data available, it has also become custom to parameterize model predictions against crystallographic B-factors and ensemble variation in NMR models. This practice does not come without assumptions, however, as neither of these are direct observations of thermal motion, and in the case of B-factors the experimental conditions do not reflect the solvent environment for which one would usually want the model to apply. Indeed the parameterization against B-factors tends to make long-range contacts stiffer than models obtained from MD-simulations and radial distribution analysis {Fuglebakk, 2013 #15}. Notably, a wide range of cut-off values (from 8 to 15 angstroms) has been used in cut-off based models to represent the interatomic interactions. Some models are sensitive to these values, but their implications on interpretation are largely left neglected. In recent years attempts have been made to carefully quantify how these assumptions affect the parameterization {Riccardi, 2010 #11;Hinsen, 2008 #12;Soheilifard, 2008 #13;Fuglebakk, 2013 #15}. As these benchmarking studies show that the performance of different ENMs depend on the benchmark chosen, researchers should carefully consider which benchmark they trust for their application, and choose or define their model accordingly. ENMs

can also be modeled to reflect a crystalline environment {Kundu, 2002 #580}, and parameterizations obtained for such models can potentially help in parameterizing single protein ENMs. Even so, exact interpretation should be made cautiously, as B-factors are heavily influenced by non-thermal contributions {Hinsen, 2008 #12;Soheilifard, 2008 #13}.

3. Validation

Early studies comparing coarse-grained NMA and experimental structural data, or molecular dynamics simulations were used to validate the method. Validation against detailed molecular mechanics force fields have shown on large protein datasets that ENMs reproduce well the slow dynamics obtained from molecular simulations (e.g.{Micheletti, 2004 #7; Pontiggia, 2007 #42;Rueda, 2007 #14; Moritsugu, 2007 #5; Yang, 2008 #41; Moritsugu, 2008 #1830;Moritsugu, 2009 #6; Skjaerven, 2011 #53; Fuglebakk, 2013 #15}).

Furthermore, several studies have focused on validation against experimental data; they evaluated the number of low-energy modes necessary to describe the structural difference between two different x-ray structures (say one opened and one closed) of the same protein using the overlap between the calculated set of modes and the structure difference vector as a quality measure. These studies show that in many cases a few low-energy normal modes account for most of the structure difference (in term of difference vector) {Marques, 1995 #960;Hinsen, 1999 #138;Tama, 2001 #1040;Krebs, 2002 #926}. Hinsen et al. {Hinsen, 1999 #36} compared domain identifications from an ENM with those obtained from internal distance differences in experimentally determined conformations of Citrate Synthase, HIV-1 Reverse Transcriptase and Aspartate Transcarbamylase. Sanejouand and coworkers systematically analyzed the agreement between low energy normal modes and small data sets of experimentally determined structures in different conformational states {Tama, 2001 #37;Delarue, 2002 #38}. Krebs et al. showed that more than half of a set of 3800 protein motions could be described by only two of the lowest energy normal modes {Krebs, 2002 #926}. Utilizing the large number of structures determined for some proteins, the structural variation can be decomposed into principal components and compared with normal modes, as done by for example Bakan and Bahar {Bakan, 2009 #39}

. In all of these studies the conformational changes of the proteins were found to be well described by the lower energy normal modes intimately linked to the protein's structure.

In addition ENMs have been used as a tool for characterization in many case studies of proteins and macromolecular complexes. In many such studies the normal mode analysis is validated by comparing with conformational change, or by testing the insights obtained by independent means {Valadie, 2003 #40;Tama, 2003 #41;Reuter, 2003 #42;Zheng, 2007 #43}. Comparison of predictions from ENMs with molecular dynamics simulations has also been used to validate and benchmark models {Micheletti, 2004 #7;Rueda, 2007 #14;Fuglebakk, 2013 #15;Moritsugu, 2007 #5;Moritsugu, 2008 #1830;Moritsugu, 2009 #6}.

4. Comparing intrinsic dynamics: getting quantitative

Comparisons of principal modes of motion have been done successfully by manual inspection and expert judgement comparing calculated properties. In recent years, we have seen progress on ways to assess the similarity of motion quantitatively. This is particularly useful for large-scale statistical analysis, benchmarking and clustering.

4.1. Similarity measures

ENMs can predict atomic fluctuation through Eq. [7], and such fluctuation profiles can be compared to fluctuations obtained from other structures or models by an appropriate association measure, such as the squared inner product, SIP:

$$\mathrm{SIP}(\mathbf{a},\mathbf{b}) = \frac{\left(\sum_{i=1}^{N} a_i b_i\right)^2}{\left(\sum_{i=1}^{N} a_i^2\right)\left(\sum_{i=1}^{N} b_i^2\right)} = \left(\frac{\mathbf{a}^{\mathrm{T}}\mathbf{b}}{||\mathbf{a}||\ ||\mathbf{b}||}\right)^2 \qquad [11]$$

where **a** and **b** are vectors of size N with elements quantifying the atomic fluctuation of each atom in the model. Correlations measures have also been commonly applied. Since exact energies are not reliably predicted by ENMs, any quantity that factors in the eigenvalues, such as the atomic fluctuations, are necessarily compared in a normalized fashion.

As mentioned above, motions calculated from ENMs are only valid for small displacements from equilibrium, and the inference to large deformations involves assuming that the interatomic couplings are relevant for longer timescales. It is therefore preferable to compare the normal modes or the covariance matrices of the

ENMs, rather than atomic fluctuations, which only indirectly reflect the covariance structure of the protein. This concern does indeed have practical implications as we reported recently {Fuglebakk, 2012 #28;Fuglebakk, 2013 #15}. For comparing sets of normal modes, the Root Mean Squared Inner Product (RMSIP) of the lowest modes has been a common choice:

$$\mathrm{RMSIP}(\mathbf{V},\mathbf{W}) = \left(\frac{1}{n}\sum_{m=1}^{n}\sum_{l=1}^{n}\left(\mathbf{v}_m^{\mathrm{T}}\mathbf{w}_l\right)^2\right)^{\frac{1}{2}} \qquad [12]$$

where **V** and **W** are sets of normal modes or principal components, and the sum runs over the nontrivial modes of lowest energy or highest variance. The constant *n* defines, somewhat arbitrarily, a subspace of protein motion that is considered accessible by low-energy motion. The RMSIP quantifies how similar the directions of this low-energy subspace are for two protein models. Since the modes are orthogonal, the RMSIP would be exactly 1 if the summation was extended to the entire set of modes. Typically, this measure has been applied with n=10, following Amadei et al. {Amadei, 1999 #29}. As the RMSIP does not represent the energetic separation between modes in the sets, measures that incorporate eigenvalues as well have been proposed. Hess {Hess, 2002 #30} defined an overlap function, OV:

$$\mathrm{OV}(\mathbf{A},\mathbf{B}) = 1 - \left(\frac{\mathrm{tr}\left(\left(\mathbf{A}^{\frac{1}{2}}-\mathbf{B}^{\frac{1}{2}}\right)^2\right)}{\mathrm{tr}(\mathbf{A})+\mathrm{tr}(\mathbf{B})}\right)^{\frac{1}{2}} \qquad [13]$$

where **A** and **B** are covariance matrices. $\mathbf{A}^{1/2}$ is the matrix that decomposes into the same orthonormal eigenvectors but with eigenvalues that are the square root of those in **A**.

Here the normalization is realized by dividing by the sum of matrix traces in the denominator. The trace of a covariance matrix is equal to the total variation in the system. Trace normalizing covariance matrices was also applied by e.g. Fuglebakk et al. {Fuglebakk, 2012 #28} who applied their similarity measures to normalized matrices $\tilde{\mathbf{C}}$:

$$\tilde{\mathbf{C}} = \frac{1}{\mathrm{tr}(\mathbf{C})}\mathbf{C} \qquad [14]$$

Applying the measure to these normalized matrices reveals an alternative making clear the similarity with the RMSIP:

$$\mathrm{OV}(\tilde{\mathbf{A}},\tilde{\mathbf{B}}) = 1 - \left(1 - \sum_{m=7}^{3N}\sum_{l=7}^{3N}(\kappa_m\mu_l)^{\frac{1}{2}}\left(\mathbf{v}_m^{\mathrm{T}}\mathbf{w}_l\right)^2\right)^{\frac{1}{2}} \quad [15]$$

where $\tilde{\mathbf{A}}$ is decomposed into eigenvectors $\mathbf{v}_m$ and eigenvalues κ_m, $\tilde{\mathbf{B}}$ into eigenvector $\mathbf{w}_l$ and μ_l and the sums run over the nontrivial modes. Note that the numerator has the form of RMSIP weighted by the eigenvalues, and that it is equal to the denominator for identical sets of eigenvectors and eigenvalues. A similar measure was also proposed by Carnevale et al. {Carnevale, 2007 #33}.

Considering the problem of comparing the intrinsic deformations in proteins as a matter of comparing their Boltzmann distribution, the field of multivariate statistics provides many measures of distance or similarity. Of those, the Bhattacharyya coefficient and the closely related Bhattacharyya distance have been adapted for comparing internal deformations of ENMs{Fuglebakk, 2012 #28}. The Bhattacharyya coefficient, BC is defined as:

$$\mathrm{BC}(p_a,p_b) = \int p_a(\mathbf{r})^{\frac{1}{2}}p_b(\mathbf{r})^{\frac{1}{2}}\,d\mathbf{r} \quad [16]$$

where p_a and p_b denote probability density functions (PDFs) for a multivariate random variable. For comparing internal deformations of proteins, the distributions can be taken to be mean-centered, and for ENMs, the PDFs will be Gaussian with the covariance matrix specified in Eq. [6]. For mean-centered Gaussian distributions with trace normalized covariance matrices, BC has the closed form:

$$\mathrm{BC}(p_a,p_b) = \frac{|\tilde{\mathbf{A}}|^{\frac{1}{4}}|\tilde{\mathbf{B}}|^{\frac{1}{4}}}{\left|\frac{1}{2}(\tilde{\mathbf{A}}+\tilde{\mathbf{B}})\right|^{\frac{1}{2}}} \quad [17]$$

where $\mathbf{A}$ is the covariance matrix of p_a and $\mathbf{B}$ is the covariance matrix of p_b, and vertical bars denote the matrix determinant. However, the measure is only defined for positive-definite covariance matrices, and an approximation to $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ has to be made due to the presence of the trivial modes. This has been solved by projecting the matrices to a lower dimensional subspace chosen from the eigendecomposition of the mean matrix in the denominator {Fuglebakk, 2012 #28}.

The information used by the different measures is illustrated in Figure X. These simple bivariate distributions can be taken to represent deformations of a molecule represented by two coordinates. The SIP ignores any directionality of motion, and simply adds up the total variance of position. The RMSIP considers the agreement of

direction for all pairs of eigenvectors that ranks among the *n* highest principal components (corresponding to lower energy normal modes). For this simplified example only an RMSIP with *n*=1 can be considered, which amounts to only comparing the principal component with the maximal variation of each distribution. Note that the comparisons with the other principal component of either distribution are represented with dotted lines for comparison with the OV. The OV compares all pairs of eigenvectors, but factors in the variance along each direction. This is illustrated by vectors with lengths proportional to the standard deviation along the principal directions, and can be contrasted with the normalized vectors considered by the RMSIP. The BC quantifies the similarity of the PDFs, which is here illustrated by the overlapping region of the two elliptical distributions.

4.2. Structural alignment

When the intrinsic motions of non-identical structures are compared, it is necessary to first obtain a description of which parts of the different structures are to be compared with each other. For example, a structural alignment can describe which amino acid residues are in structural correspondence to each other between two or more structures. Comparing distant homologues provides a challenge in defining what parts of the proteins to compare. This is commonly solved by structural alignment, which is a challenging problem, particularly for the simultaneous alignment of sets of proteins {Berbalk, 2009 #69;Hasegawa, 2009 #76;Marti-Renom, 2009 #89;Russell, 1992 #126}. A thorough discussion about defining comparable regions of a protein and on some strategies for aligning pairs of proteins using models of their intrinsic flexibility is described in a review of Micheletti {Micheletti, 2012 #31}.

When considering an alignment for comparing multiple structures, sequence identity and volumetric differences tend to pose a big challenge to finding equivalent atomic coordinates between them. The optimum solution between two structures, let alone many, tends to scale with sequence length and variability. Moreover, there is also the question of the most reasonable way of assessing a resulting solution, even though root mean squared deviation (RMSD) is generally accepted as the standard across different tools within the field. Due to the dynamic nature of structures, many alignment solutions involve a component of flexibility to achieve a better fit between structures. Yet these solutions are mostly available for pair-wise alignments. Most

multiple structural alignment methods involve computing all pair-wise alignments between a set of structures, before producing a consensus between all of them{Marti-Renom, 2009 #103}. The differences between multiple structure alignment programmes involve the choice of geometric reference points, such as secondary structure or Cα atoms, algorithm for aligning them in a pairwise fashion or identifying a consensus core alignment to optimise iteratively, and the way these are scored at the end. Problems unique to multiple structure alignment involve the length of consensus alignment between multiple structures, and pairwise RMSDs within the set {Ma, 2014 #104}. We find that in order to compare structures effectively, it is essential to have a robust alignment that is able to take into account natural and yet unique variations within a set of proteins. Previously, we have found that the results of comparative analysis are sensitive to the quality of the alignment {Fuglebakk, 2012 #28}, especially if the set contains structures that are related at the SCOP family and superfamily levels {Murzin, 1995 #105}.

To illustrate this, we constructed multiple structural alignments for a large set of proteins with triosephosphate isomerase-like domains (c.1 TIM alpha/beta barrel dataset from Fuglebakk et al {Fuglebakk, 2012 #28}) using two popular programs, STAMP{Russell, 1992 #106} and MUSTANG{Konagurthu, 2006 #107}. The set consists of structures from four families, from two different superfamilies. The triosephosphate isomerase possesses a fold that is tricky to align, as it is completely symmetrical in its enclosed barrel-like configuration that consists of 8 strands and 8 helices. As such, it is a challenge for even the most sophisticated algorithms to align, especially when the sequence identity is low, due to the abundance and diversity of this fold {Nagano, 2002 #112}. Visual inspection of the superimpositions provided by STAMP shows that it is heavily biased towards the N-terminus, where the alignment is optimized, losing symmetry in the points of common reference towards the C-terminus (Figure 1). On the other hand, the MUSTANG alignment is able to provide a superimposition that is less sparse, with regards to these points in common, and with these are distributed all over the structures (Figure 2).

STAMP {Russell, 1992 #106} relies on an algorithm that assesses pairwise alignments within a set of protein and extends it in a progressive manner with the aid of hierarchical clustering. In this process, there is a chance that an error made in introducing a gap earlier is propagated to the final output. MUSTANG {Konagurthu,

2006 #107} employs a similar algorithm that differs in the way that it refines the alignment once the pairwise alignments are completed by introducing an intermediate step where the residue position equivalences are assessed globally (in the context of the other proteins in the set).

The effect of the difference in quality between these alignments propagates when we try to cluster these structures to their family and superfamily levels based on their intrinsic dynamics, using the Bhattacharya coefficient score {Fuglebakk, 2012 #28}. The difference is striking: we see some separation between the structures at the family and superfamily level which is not totally identical to the SCOP annotation, but performs better than fairly mixed clustering that we see resulting from STAMP. Since we find a large overlap between structural similarity and dynamics in most cases, we find that using an appropriate tool to align multiple structures is important in having reliable results when comparing their intrinsic dynamics.

4.3. Comparing only the conserved regions

Once the issue of determining the corresponding parts of each protein within a set is resolved, we can proceed to define a subset or a core between all of these proteins that can be compared dynamically. In such a procedure, each protein can be partitioned into a subset of core atoms A, and a subset of excluded peripheral atoms B. While only A has a corresponding part in all the structures compared, B is still linked to the dynamics of A and should be retained in the calculation of motion to preserve its influence. This needs to be observed when measures are normalized. In the case of comparing normal modes or covariance matrices, lower dimensional matrices describing the motion of only A needs to be obtained. Many deformations of the proteins can be consistent with A acting as a rigid body, while B is seen to deform internally. Since B is defined to not be comparable between structures, it is desirable to express the internal deformations of A in a way that is consistent with how it deforms in the context provided by B. Mathematically, the problem is manifested by the fact that the parts of the eigenvectors corresponding to only A is not generally orthogonal. One common way to deal with this problem is to define a potential for A, which is restrained by the presence of B {Hinsen, 2000 #35;Carnevale, 2007 #33}. Assuming that B deforms along the direction of minimal energy, such a restrained potential can be obtained by differentiating Eq. [3] with respect to deformations of B.

Substituting these minimal energy deformations of B back into Eq. [3] gives the Hessian of the constrained potential:

$$\tilde{\mathbf{H}} = \mathbf{H}_{aa} - \mathbf{H}_{ab}\mathbf{H}_{bb}^{-1}\mathbf{H}_{ab}^{\mathrm{T}} \qquad [18]$$

where the Hessian of the full potential is partitioned so that $\mathbf{H}_{aa}$ reflect interactions in A, $\mathbf{H}_{ab}$ reflect interactions between A and B, and $\mathbf{H}_{bb}$ reflect interactions in B. This method was originally introduced for ENMs, but now has also been extended to all-atom and hybrid quantum mechanics/molecular mechanics potentials, for its recognized potential as an analysis method {Woodcock, 2008 #95}.

When normal modes or covariance matrices are expressed in rotational variant coordinate systems, like the formalism in Cartesian coordinates described above, one needs to make sure that they are expressed in a identical or similar rotations. When comparing models of identical proteins, this can be solved exactly. In cases where different proteins are compared, an approximation to a common rotation is typically obtained by rotating to minimize the sum of squared distances between aligned residues. For comparing proteins with very different equilibrium structures, the validity of such an approximation to a common reference frame might become a concern. Possible solutions include considering internal coordinates {Mendez, 2010 #8} or comparing rotationally invariant properties of the normal modes, like the correlation matrix in Eq. [8].

5. Strategies and Applications of Comparative Analysis

Comparing multiple structures is a natural extension to the study of intrinsic dynamics. In the case of ENMs, where it is computationally efficient and inexpensive, it's been seen as a logical choice for analysis of large sets of structures in many ways. The modes vectors produced from ENMs are informative on their own, and provide a good qualitative description of a protein's inherent flexibility. Luo and Bruice did one example of such a qualitative analysis, where they conducted a visual inspection of the normal modes vectors on six structures along the dihydrofolate reductase reaction mechanism {Luo, 2009 #155}. With this analysis, they focussed on the regions (such as the M20 loop and sub domain rotations) that were seen to undergo conformational changes during catalysis, and found that they were consistent with the principal motions from the low-energy modes, and were able to further validate these findings against NMR, kinetic and molecular dynamics studies found in literature.

In general, we find that such analyses benefit from quantification of dynamical properties, and here, we outline a selected list of examples where ENMs have been used to compare the dynamics of multiple structures, whether from the same sequence or not, related by homology or fold. Some of the early examples discussed have been used to validate the ENM as a viable model for intrinsic dynamics analysis, yet the strategies employed are applicable to most comparative applications. In most of the examples, atomic fluctuation profiles (Eq [7]) were used to compare between structures and against experimental B-factors, while the use of the overlap analysis (such as the squared overlap, Eq. [9]) was also very common. In addition, we find that comparing covariance/correlation matrices, using similarity measures (such as the RMSIP, Eq. [12]) and perturbation response methods {Zheng, 2005 #55} are also useful techniques when comparing dynamics.

5.1 Comparing multiple structures of the same protein sequence: conformational changes

The squared overlap between modes and the structural difference from one conformation to another has been introduced as a way to understand the transformation between two states of an enzyme (Eq [9]). This analysis allows for the identification of modes that contribute to the conformational change seen {Marques, 1995 #49;Tama, 2001 #37; Reuter, 2003 #42}. Traditional dynamics studies that compare two extreme states, e.g. fully open ligand-free conformations vs. fully closed ligand-bound conformations, lead to interpretation that the modes with high overlap with the difference between conformations are the ones important for conformational change. These modes tend to be interpreted as the transition path between active and inactive states. In general, mapping conformation transition paths are a much more complex affair that require more detailed and rigorous calculations than the overlap analysis to estimate {Maragakis, 2005 #118;Whitford, 2008 #90;Togashi, 2010 #100}.

Even so, the method has been useful in understanding the changes in flexibility between states. Extending pair-wise comparisons can come in the form of performing serialised overlaps between multiple pairs of structures of different conformational intermediates, or a large-scale survey of conformational transitions. An example of a large-scale survey of transitions is the work from Stein and colleagues who performed

serial overlap analyses of pairs unbound to bound conformations (multiple pairs in some cases, where more than one ligand-bound conformation was available for an enzyme) from a total of over 12,000 structures {Stein, 2011 #156}. This was one of the analyses used to assess the cost of conformational change upon ligand binding, and whether they fit the lock-key, induced-fit or conformational selection binding models.

ENMs can be used to produce covariance matrices (Eq. [6]) that can be compared between them to understand the difference in dynamics for different states. The work of Seckler et al.{Seckler, 2013 #83} is one such example, where the authors use ENMs in addition to structural comparison; the authors retrieved 52 structures of HIV-1 reverse transcriptase, and compared them to reference structures using a measure of dynamics similarity called the *covariance complement* (a form of the OV measure, Eq. [13]) . The structures differed in state such that some were ligand-free while others had DNA, RNA, adenosine triphosphate (ATP) and various inhibitors bound. Further they found linear variation of RMSD with the covariance complement to be a signature of functional state, and showed that the ratio between the two measures can be used to cluster these 52 structures into three main levels. These levels corresponded to their level of activity based on the ligand-types. This is an example where dynamics is used to distinguish between the effects of ligands.

Allosteric effects of ligands and their ability to cause changes in flexibility have often been explored using ENMs {Tehver, 2009 #58;Motlagh, 2014 #77}. While the allosteric effect is commonly known to cause a large conformational change in a structure, Rodgers et al. explored the hypothesis that the low-frequency normal modes are able to propagate allosteric signals without causing a large conformational change in a family of transcription factors called CRP/FNR {Rodgers, 2013 #57}. They constructed ENMs for structures of the Catabolite Activator Protein (CAP) from *Escherichia coli* representing unliganded, single and double-liganded forms and introduced mutations outside the substrate-binding pocket by varying the spring constants of all springs attached to a single residue. This was a strategy for probing changes to the free energy of substrate binding. They found that the regions that experienced the greatest change in cooperativity were not necessarily adjacent to the substrate-binding site. They used their method to predict residues involved in

allosteric signalling in CAP and validated their findings on a homologue GlxR, through a combination of ENM, MD and experimental results.

The large increase in the number of X-ray structures has led to the opportunity of analysing preferred conformations with multivariate statistical analyses such as Principal Component Analysis (PCA). PCA has the advantage of reducing the dimensionality of large-scale data into basis vectors (or principal components), ordered based on how much of the structural variance they describe. The principle components in this case describe the direction of the structural variance in the dataset (due to experimental conditions or evolutionary relationships) rather than the thermal forces in ENM or MD calculations. One strategy employing PCA involves calculating principal components for a large collection of structures for a given protein, and comparing the resulting principle components to modes obtained from ENM calculations on representative structures. Yang and colleagues have shown that normal modes can directly be compared with principle components extracted from a large set of structures from HIV-1 protease, providing a direct comparison of calculated values to experimental data {Yang, 2008 #51}. Both Bakan and Bahar {Bakan, 2009 #39}and Katebi et al. {Katebi, 2014 #52} applied this strategy to different enzymes; structures of HIV-1 reverse transcriptase, p38 MAP kinase and cyclin-dependent kinase 2 with and without inhibitors were analysed by Bakan and Bahar and all available structures of HIV triosephosphate isomerase (TIM) by Katebi et al. They were able to relate the variation in the structural space to intrinsic dynamics and further to function.

5.2 Comparing dynamics between different oligomeric/multimeric states

The variations in the dynamics of monomers often translate to changes in global structure, whether they are perturbed by mutations or ligands. As a recommendation, when comparing oligomeric structures, one should be wary of the effect of calculating the modes of proteins in different oligomeric states before drawing conclusion on the dynamics of the system or making a one-to-one comparison. The lowest modes are different from one oligomeric state to the other, where subunit-subunit motions make up the lowest energy modes in multimeric assemblies. Eigenvectors calculated on entire structures with different oligomeric states are thus not necessarily comparable. Since monomers usually possess differences in conformation based on their

oligomeric state, comparing monomers extracted from different states provides sufficient basis for observing changes in their dynamics {Marcos, 2011 #97}.

Another notable example of working with multiple subunits is in the study of Alzheimers' Aß(1-40) amyloid fibrils, where Xu and colleagues constructed models of these protein assemblies based on two forms of naturally occurring symmetries and varying lengths of the fibrils{Xu, 2010 #59}. They were able to characterise the effects of the fibril size on the overall flexibility of these large structures, and changes to the low-energy motions. Similarly, Polles et al explored the flexibility in different assemblies of a heterogeneous set of virus capsids based on fluctuations-based analyses and domain decomposition {Polles, 2013 #137}. Other studies dedicated to the comparison of dynamics to changes in oligomeric state have been reported for monomeric, dimeric and tetrameric states of GPCRs {Niv, 2008 #115}, dimeric and hexameric (trimer-of-dimer) states of the serine receptor Tsr{Hall, 2012 #111}, and monomeric and dimeric states of the p53 protein{Kantarci, 2006 #109}.

5.3 Comparing dynamics between more distantly related proteins

Structure comparison has been long established as a means to understanding the evolution of proteins, as has the conservation between sequence and structure since the work of Lesk and Chothia {Chothia, 1986 #110}. In many cases, even from visual inspection, a structure is not just seen to encode dynamic information, but also a historical time-point in the evolution of the family to with it belongs. Some sequence mutations, insertions and deletions can be accommodated by the plastic deformations of a common architectural core and retain the precise geometry of the active site, even if peripheral regions or accessory domains vary {Hasegawa, 2009 #118}. Hence it has been of great interest in linking observable structural and sequence evolution in the conservation of dynamics, especially in a family of proteins.

The observation that low energy normal modes so frequently appear in functional motion {Krebs, 2002 #92;Nicolay, 2006 #16} has motivated investigations into their evolutionary conservation across protein families. The globins have served as a good example of a well conserved yet diverged group of proteins and has been a subject of quantitative comparisons by Maguid and colleagues {Maguid, 2005 #17}. They developed a method to quantify similarities between the collective modes using a singular value decomposition approach to find representative vectors to describe the

dynamics of an aligned core of structures. This work laid the foundation for exploring the evolutionary conservation of dynamics in the lowest energy modes, which they further developed and validated across larger standard datasets from the family to superfamily levels {Maguid, 2006 #18;Maguid, 2008 #19}. While this work clearly confirms that low energy normal modes are conserved between structurally conserved proteins, it is interesting to note that this conservation can be explained as the structural response to random perturbations, rather than necessarily selective pressure on certain kinds of motion {Echave, 2008 #24;Echave, 2010 #25;Echave, 2012 #26}.

In 2007, Carnevale et al {Carnevale, 2007 #33} introduced the idea of performing pair-wise alignments and comparing dynamics for the regions conserved by giving each pair an overall score such as RMSIP (Eq. [11]), which they used on pairs of proteases with very low sequence identity. Partial pair-wise alignments between the proteases allowed them to conclude that often the dynamical conservation far exceeded the structural conservation.

Others have explored the idea that the effect of sequence changes throughout the evolution of a protein structure would be along its principle modes. Work by Leo-Macias et al {Leo-Macias, 2005 #20; Leo-Macias, 2005 #21} performed large scale deformation analysis on the multiple alignments of the cores of 35 protein families and compared them against their evolutionary deformations gained from the PCA these families via the RMSIP score. They were able to able to relate regions with the greatest evolutionary variability with regions that experience greater thermal fluctuations. This work was further validated with comparisons against MD simulations{Velazquez-Muriel, 2009 #22} and further reinforced by work done on the Ras GTPase superfamily {Raimondi, 2010 #23}.

In a more recent study combining allostery and evolutionary conservation, Kolan et al. report the role of the lowest energy modes on the mechanical motions and conformational changes of six members of the GPCR family {Kolan, 2014 #119}. When comparing between the ligand-bound and ligand-free states, they found that the slowest modes are well suited to describe components of the activation mechanism. They compared the overlap of their slowest modes by calculating the normalized mean squared displacements of the aligned Cα atoms as a correlation score and showed that all the GPCR members except rhodopsin agree well, and concluded that

rhodopsin was not representative of all GPCRs in its motions. They concluded that the ENM calculations were able to capture the long-range mechanism of GPCR activation, where binding in the extracellular domain can cause a conformational change in the cytoplasmic domain.

Another example of linking the conservation of certain structural/sequence motifs to function is displayed by the work of Lukman and Grant{Lukman, 2009 #79}. They surveyed maltose transporters and characterised a network of residues that have an influence on the overall dynamics of the proteins in different conformational states. This work is an example of analysis inspired by the developments in the perturbation-response analysis by {Zheng, 2005 #55}, who had earlier studied the conservation of dynamics in distantly-related motor proteins by comparing the conformational changes experienced by myosin, F1-F0 ATPase and kinesin {Zheng, 2003 #96}. In this case, the authors concluded that while the large conformational change seen in myosin and F1-F0 ATPase was consistent with motions that can be described as a power-stroke type movement, while the kinesin followed a Brownian ratchet-type mechanism. The perturbation-response method has gained greater traction as seen in the efforts to develop a useful metric {Nevin Gerek, 2013 #121; Atilgan, 2009 #120} to describe a single residue's response to an applied force in a given position, as a predictive tool.

Studies like this have also prompted initiatives to categorise protein structures dynamically {Warren, 2014 #89}. Further databases storing results from normal mode analysis using ENMs on large number of structures have been built, such as ProMode Elastic{Wako, 2011 #121} or MolmovDB (Database of Macromolecular Movement{Gerstein, 1998 #128}. These show the interest of the community and the potential of ENMs for the characterisation of intrinsic dynamics in a way that can complement existing structural classifications systems. In addition, there have been efforts in using dynamic information as a means of alignment different proteins, and their developments have provided insight into comparing dynamics in general {Carnevale, 2007 #33; Davis, 2014 #98;Tobi, 2012 #93}.

Studies like this have also prompted initiatives to categorise protein structures dynamically {Warren, 2014 #89}. Further databases storing results from normal mode analysis using ENMs on large number of structures have been built, such as ProMode

Elastic{Wako, 2011 #121} or MolmovDB (Database of Macromolecular Movement{Gerstein, 1998 #128}. These show the interest of the community and the potential of ENMs for the characterisation of intrinsic dynamics in a way that can complement existing structural classifications systems. In addition, there have been efforts in using dynamic information as a means of aligning different proteins, and their developments have provided insight into comparing dynamics in general {Carnevale, 2007 #33; Davis, 2014 #98;Tobi, 2012 #93}.

5.4 Comparing structures with different folds

In the paradigm where the conservation of dynamics is due to structural similarity and not vice versa, the comparison of dynamics based on shape and fold, independent of sequence similarity or conservation, has also been a topic of great interest {Lu, #116;Tama, 2006 #65}. Since proper folding of the protein is a requirement for function in many cases, it is natural to seek to understand how the fold affects function, and their principal modes of motion is an important ingredient in understanding functional properties of the fold.

Hollup et al. showed that computer-generated models based on ideal structures, stripped of influences of sequence conservation and evolutionary links, could be used reliably in the analysis of dynamics {Hollup, 2011 #22}. They showed that the spatial arrangement of secondary structures in a protein is an important component of the low energy modes, while the loops connecting these elements play a minor role. Another study characterised the motions of two proteins with cylindrical symmetry, the beta-barrel Dronpa and the toroidal DNA-clamp, as ideal structures and compared them both qualitatively to find similarities in their global motions {Hu, 2012 #64}.

6. Computational tools and frameworks

The simplicity of the ENMs makes them relatively easy to implement if routines for the necessary linear algebra is provided. This makes it easy to integrate ENMs with other kinds of structural analysis. Most model developers also make implementations available online or upon request. In addition the interested user can choose from a range of tools and frameworks available for computing and analyzing ENMs. The Molecular Modelling Toolkit {Hinsen, 2000 #452} and ProDy {Bakan, 2011 #1638} are libraries for the programming language Python that support normal mode decomposition, analysis and visualization of ENMs. For the statistical computing

software R, ENMs are integrated into packages for analyzing Molecular Dynamics data like LOOS {Romo, 2009 #46} and Bio3D {Grant, 2006 #45}. ΔΔPT {Rodgers, 2013 #44} is a collection of scripts for ENM and principal component analysis that allows the application of a range of ENMs without requiring familiarity with programming. A range of web servers are also available such as WEBnma{Hollup, 2005 #472}, ElNemo{Suhre, 2004 #971}, ANM webserver {Eyal, 2006 #282}, KOSMOS{Seo, 2012 #907}, NMSim {Kruger, 2012 #571}, NOMAD-Ref {Lindahl, 2006 #633}. These provide a variety of analysis on ENM normal modes, typically making the analysis accessible for an audience less experienced with computational analysis. Another initiative that aims to take ENMs to a wider audience is the software Maven, provided as a standalone application for analysis and visualization of ENMs {Zimmermann, 2011 #47}.

Of the web servers, WEBnma is the only one currently supporting comparative analysis with the use of structural alignment information along with the submitted structures. It also provides an easy access to the Bhattacharyya Coefficient score. ProDy allows for the comparison of sequence evolution data (with the implementation of a co-evolutionary analysis tool Evol) and intrinsic dynamic information from ENMs {Bakan, 2014 #144}. The latest releases of Bio3D provide an implementation that provides the framework for "automated ensemble analysis methods", which includes multiple sequence alignment and a selection of similarity measures and correlations analysis. Amongst the algorithms proposed for dynamics-based alignment of proteins, one is also made available as a web-server.

7. Conclusion/Perspectives

The use of ENMs for comparative analysis of proteins dynamics has lead to greater understanding in the conservation of dynamics across structures with different conformations and within a proteins family. Moreover, there have been more and more evidence that comparing dynamics is a viable way for gaining greater understanding for the mechanisms employed by proteins for their function. Efforts have been made lately to evaluate the effect of the choice of similarity measures, the ENMs parameter and the structural alignments. The results of these studies,

summarized in this manuscript will be useful for users getting started on comparing the dynamics of proteins in a wide variety of settings.

As any models, the ENMs naturally have their limits and their application for comparative dynamics analysis certainly has too. It is therefore important that users are aware of the potential impact of ENMs parameterization, structure superimposition and choice of similarity measure. We advocate for detailed reporting of the corresponding computational strategies in scientific publications. This will also contribute to increasing the credibility of this field, which alike molecular modeling in the broad sense can at times be the object of skepticism.

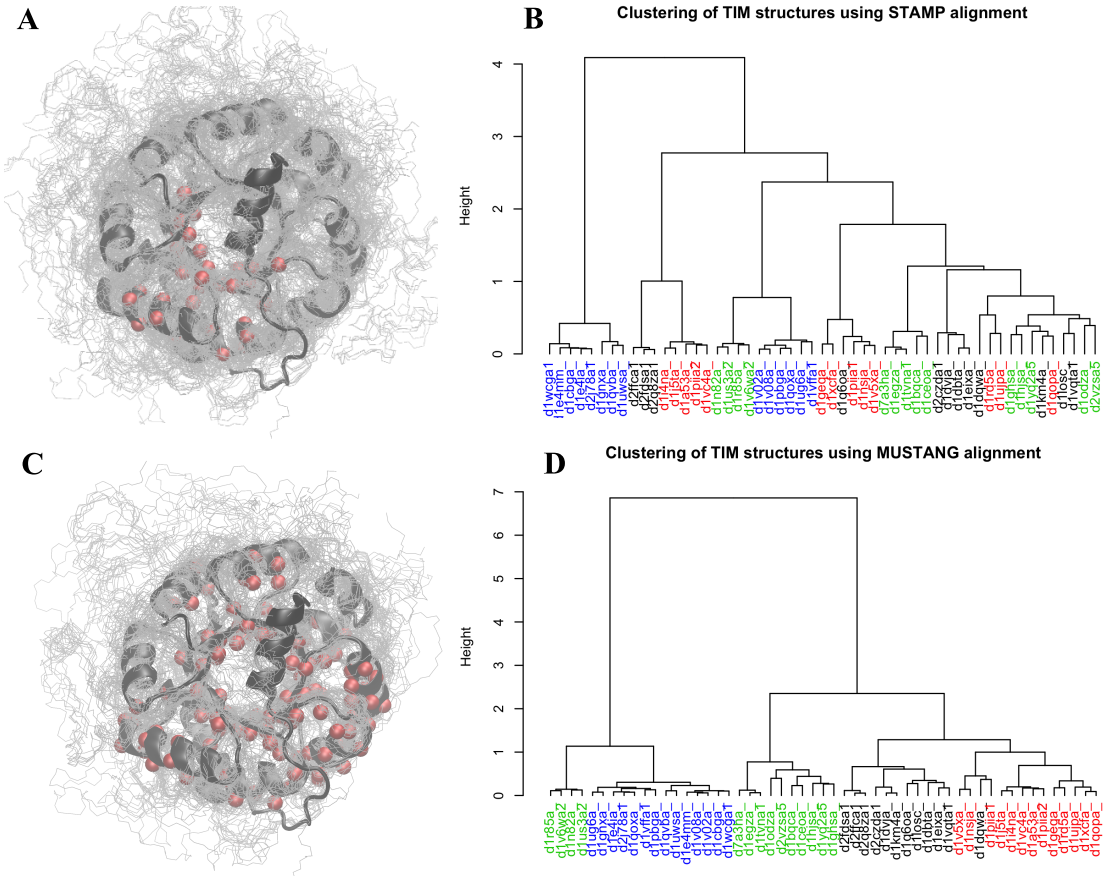
Figure captions

Figure 1:

Figure 2: Influence of the alignment methods of 53 structures with the TIM-barrel fold on similarity measures. STAMP (A) and MUSTANG (C) alignment of 53 structures with the TIM-barrel fold. The light grey lines show the superimposition of the structures, while the dark grey cartoon of the secondary structure shows one of them as a representative. The red spheres highlight the points on the structure that are conserved throughout the alignment with a bias towards the N-terminus. B) K-means clustering (with k = 4) of the Bhattacharya score analysis comparing the covariances using the STAMP (B) and MUSTANG (D) alignments. The colours signify structures that are from the same family, and are grouped such that red and black are from one superfamily, and blue and green are from the other.We see a heterogenous clustering of the structures across superfamilies and families with STAMP while we see good groupings with respect to the family level, but a less distinct separation at the superfamily level with respect to the green group.

Figure 1

Figure 2


A
B
Clustering of TIM structures using STAMP alignment
Height
C
D
Clustering of TIM structures using MUSTANG alignment
Height

Highlights