Using Elastic Network Models to Compare the Intrinsic Dynamics of Multiple Protein Structures

[1,2], [1,2], [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. Structure vs. sequence conservation, lately more and more studies investigating the conservation of dynamics in proteins….

Scope of Review. Use of elastic network models for the comparison of the intrinsic dynamics of sets of different protein structures

Major Conclusions. 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 ….

General Significance. Cf article Sarah? Protein design?

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

ABBREVIATIONS

ENM: elastic network model

1. Introduction

Folded proteins are remarkably dense but with a heterogeneously distributed density, reflecting 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.

Edvin Fuglebakk 15/6/14 18:01
Comment [1]: Ref. Used Fisher 2004 (Protein Science) in my thesis, might use that. Kuntz & Kauzmann, 1974; Squire & Himmel, 1979 seem to be classical references, but I don't have access to either.

Nathalie Reuter 15/6/14 18:01
Comment [2]: Use this? P. J. Fleming and F. M. Richards, "Protein packing: dependence on protein size, secondary structure and amino acid composition," *J. Mol. Biol.*, vol. 299, no. 2, pp. 487–498, Jun. 2000.

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, where she showed that a single-parameter potential could reproduce the slow dynamics obtained with a more complicated potential {Tirion, 1996 #42}. This simplication 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 energy minimization associated with 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 normal mode analysis (NMA) of proteins{Brooks, 1985 #7;Go, 1983 #8;Levitt, 1985 #9;McCammon, 1977 #6;Noguti, 1982 #10;Brooks, 1983 #5}.

Sandhya Tiwari 16/6/14 13:20
Deleted: The field was pioneered by Monique Tirion

Sandhya Tiwari 16/6/14 13:19
Deleted: who showed in 1996

The robustness of NMA with ENMs for the description of protein's slow collective motions has almost come as a surprise to many. 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 have shown on large protein datasets 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}). Further several studies have shown that in many cases a few low-frequency 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-frequency 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.

In a comprehensive review article published in 2013, Cristian Micheletti summarizes the developments and applications of methods for comparing proteins internal dynamics{Micheletti, 2013 #44}. 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 (REF), evaluating conservation of dynamics within a homologous protein family (REF), or within a set of proteins that possess the same fold despite low sequence identity (REF). It has been shown, comparing structures of homologous proteins and their intrinsic dynamics, that protein structures evolve along the low frequency modes. A number of studies have shown that the low frequency modes are robust to sequence variations…. MORE EXAMPLES . The use of ENMs for comparative protein dynamics has the potential to teach us more about …. blah blah blah… Protein design?

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. ENMs are a model of choice for such studies even if computer power has admittedly become more affordable than it was at the advent of ENMs, and molecular dynamics

Nathalie Reuter 15/6/14 17:56
Comment [3]: Should we somewhere have a paragraph on relevance for e.g. understanding sequence-structure relationship, protein design, …
I'm not quite sure where to place this (conclusion is an ok place but it would be nice to mention it early on)

Edvin Fuglebakk 16/6/14 09:04
Comment [4]: I think it is worth clarifying the scope of Michelettis review a bit, to make it clear that our review is not redundant.

Nathalie Reuter 15/6/14 17:56
Comment [5]: Should we mention the scarcity of experimental data on protein dynamics?

Edvin Fuglebakk 16/6/14 09:07
Comment [6]: Yes, I think it is worth highligting that we are looking at conservation of a kind of motion which is not easily detected by experimental means.

simulations on microseconds time-scale are becoming the rule rather than the exception. 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 such as the dynamics property to compare, the tool for sequence or structure alignment, the choice of similarity measures, etc…

Edvin Fuglebakk 16/6/14 14:22
Comment [7]: Is this a bit strong ?

After an introduction to ENMs, we will … comparative analysis on the decomposition of motion that can be obtained by ENMs or similar techniques. In particular,….

2. Elastic Network Models

2.1. Formalism

Since Tirions contribution {Tirion, 1996 #1}, further simplifications of the ENMs have been made. Realizing 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 C-alpha atoms. Atilgan et al. {Atilgan, 2001 #3} also applied Tirions model at the C-alpha granularity. Another popular density based model has been the early Gaussian Network Model {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 is therefore different from the ENMs. …

Edvin Fuglebakk 14/6/14 01:25
Comment [8]: Mention also Hwangs stuff.

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}, 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} and less granular

Edvin Fuglebakk 14/6/14 01:25
Comment [9]: Others?

representation of the proteins. Despite all this variety the ENMs can be understood in terms of a single unifying formalism, which will be detailed in the following.

The elastic network models model the protein as a network of Hookean springs connecting all residues, which are typically represented by nodes located at the center of their C-alpha 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:

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

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:

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

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:

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

Since H is a symmetric matrix the potential energy of a configuration $\boldsymbol{r}$ can be written in terms of its eigendecomposition:

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

Where $\boldsymbol{v}_m$ 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

Edvin Fuglebakk 14/6/14 01:25
Comment [10]: Tama rigid block representation.

Edvin Fuglebakk 14/6/14 01:25
Comment [11]: Amino-acid specific potentials, GO-potentials ?, MD-fitted force-constants.

Edvin Fuglebakk 16/6/14 09:12
Deleted: mass of

to the potential energy of **r**, with λ_m quantifying the energy of deforming the network in the direction of $\boldsymbol{v}_m$. These independent modes of deformation are referred to as the normal modes of the network. 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. 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 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 thermal fluctuation of node *i*. As it is often convenient to obtain a scalar quantification of the correlation of two nodes, a correlation matrix is commonly calculated, following Ichiye and Karplus {Ichiye, 1991 #9}:

$$P_{ij} = \frac{\mathrm{tr}(\mathbf{C}_{ij})}{(\mathrm{tr}(\mathbf{C}_{ii})\mathrm{tr}(\mathbf{C}_{jj}))^{\frac{1}{2}}} \qquad [7]$$

where tr denotes the trace or diagonal sum of the matrix. 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* d *j*.

As normal mode analysis has a long tradition in chemistry for analyzing 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 frequency modes {Hinsen, 2008 #34} and care should be taken not to interpret the frequencies too literally.

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. 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.

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

Sandhya Tiwari 15/6/14 18:11

Comment [12]: *I am not sure where it would be appropriate to put in concrete examples of the issue of cut-offs for example.*

For example, a wide range of values (8 to 15 angstroms) are chosen in cutoff based models to represent the interatomic interactions. Since some models can insensitive to these values, their implications on interpretation are largely left neglected.

Edvin Fuglebakk 15/6/14 18:11

Comment [13]: List some refs ?

Sandhya Tiwari 15/6/14 18:11

Comment [14]: More to be added

Sandhya Tiwari 15/6/14 18:11

Comment [15]: Funny, Bahar has a 2007 paper on how anistropic displacement factors should be an alternative to B-factors but are not currently reported: http://bioinformatics.oxfordjournals.org/content/23/13/i175.full
With the following cited as the examples you would like here:
Eyal E,et al. Anisotropic Network Model: systematic evaluation and a new web interface. Bioinformatics 2006;22:2619-2627.
Yang LW,et al. oGNM: online computation of structural dynamics using the gaussian network model. Nucleic Acids Res 2006;34:W24-W31.

distribution analysis {Fuglebakk, 2013 #15}. Notably a wide range of cut-off values (from 8 to 15 angstroms) have 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 therefore been made to more 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

Nathalie Reuter 16/6/14 01:45
Comment [16]: This paragraph has become a bit short now that I have separated parameterization and validation. I think that we should either extend on validation or we place this paragraph in the introduction

Early studies comparing NMA and experimental structural data, or molecular dynamics simulations (e.g. {Pontiggia, 2007 #978}), did validate the use of NMA with coarse-grained model. 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; Yang, 2008 #41;Fuglebakk, 2013 #15;Moritsugu, 2007 #5;Moritsugu, 2008 #1830;Moritsugu, 2009 #6; Skjaerven, 2010 #43}).

Edvin Fuglebakk 16/6/14 09:17
Comment [17]: This is not really an early example, is it?

Further several studies have focused on validation against experimental data; they evaluated the number of low-frequency 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-frequency 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

Edvin Fuglebakk 16/6/14 09:19
Comment [18]: I think we should stick to low-energy, since the formalism above introduces energetic modes.

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 frequency 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}.

Edvin Fuglebakk 14/6/14 01:27
Comment [19]: More examples of case studies ?

…

Edvin Fuglebakk 14/6/14 01:27
Comment [20]: Mention studies that show that amplitides changes faster than direction of motion in MD simulations.

4. Comparing intrinsic dynamics: getting quantitative

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

Edvin Fuglebakk 15/6/14 18:38
Comment [21]: hopefully some refs here

Edvin Fuglebakk 15/6/14 18:38
Comment [22]: refs: Fuglebakk 2012 and 2013, Grossfield papers, Micheletti papers

…

Edvin Fuglebakk 15/6/14 18:38
Comment [23]: This more quantiative approach is important when large sets of data is considered. For instance: Quantifying conservation of modes (Echave), other kinds of function strucutre alingment (Micheletti, Biggin, Bahar). Aligning evolutionary space with internal dynamics (spannish people, oroscu). Validation and Benchmarking (Grossfield, Hinsen, Mode concentration stuff). Most of this should probably be summarized in 6.

4.1. Similarity measures

As mentioned in the introduction, the motions calculated from ENMs are only valid for infinitesimal displacement 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 served the community well for a long time. Typically, this measure has been applied to the ten lowest modes 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 has been proposed. Hess {Hess, 2002 #30} and Fuglebakk et al. {Fuglebakk, 2012 #28} are notable examples in this regard.

Edvin Fuglebakk 16/6/14 09:24
Comment [24]: I am not sure we mention this in the introduction anymore, but hat is OK, it is implicit in both the introduction and the formalism, and we can explain it more explicitly here.

Sandhya Tiwari 15/6/14 18:38
Comment [25]: the covariance complement measure?

Edvin Fuglebakk 15/6/14 18:38
Comment [26]: Could elaborate on measures

…

Edvin Fuglebakk 15/6/14 18:38
Comment [27]: While a quantification of the similarity of sets of normal modes is very useful for benchmarking, model development, clustering and such applications, it is often of interest to locate structural causes underlying differences in principal degrees of freedom between proteins. Mention Julian and Atilgans perturpational analysis, mention Michelettis RMSIP decomposition, etc.

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.

Edvin Fuglebakk 15/6/14 18:38
Comment [28]: should dig out some reference for this

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 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 C-alpha 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, especially if the set contains structures that are related at the SCOP family and superfamily levels{Murzin, 1995 #105}.

Edvin Fuglebakk 16/6/14 09:29
Comment [29]: cite Bioinformatics paper, and mention any unpublished experience ?

To illustrate this, we constructed multiple sequence alignments for a SCOP superfamily of triosephosphate isomerase proteins using two popular programmes, STAMP{Russell, 1992 #106} and MUSTANG{Konagurthu, 2006 #107}. The triosephosphate isomerase possesses a fold that is tricky to align, as it a 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 of this fold in ~40% of all known protein structures. 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). On the other hand, the MUSTANG alignment is able to provide a superimposition that is sparser, with regards to these points in common, but these are distributed all over the structures (Figure).

Sandhya Tiwari 15/6/14 18:38
Comment [30]: Need to check this number again

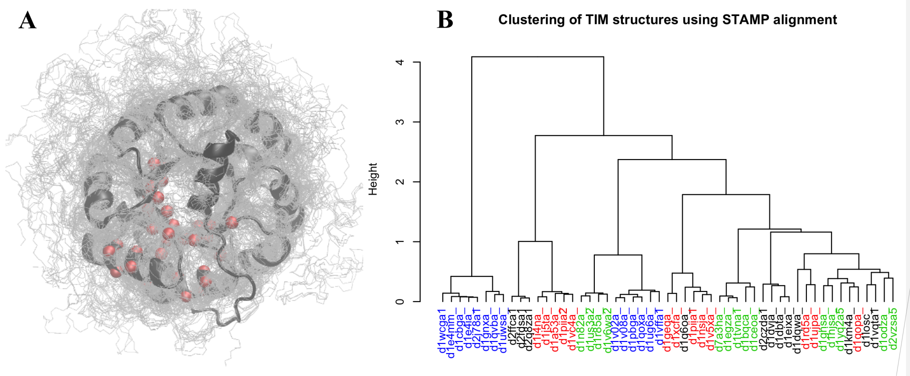


Figure 1. A) STAMP alignment of 53 structures with the TIM-barrel fold. The light grey lines show the superimposition of the structures, while the dark grey cartoon of the secondary structure shows one of them as a representative. The red spheres highlight the points on the structure that are conserved 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 alignment. 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.

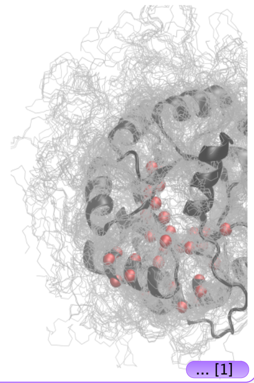
STAMP relies on an algorithm that assesses pairwise alignments within a set of protein and extends it in a progressive manner with the aid of hierarchical clustering. In this process, there is a chance that an error made in introducing a gap earlier is propagated to the final output. MUSTANG employs 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 scoring{Fuglebakk, 2012 #28}. The difference is striking: we see some separation between the structures at the family and superfamily level is not ideal 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.

Unknown
Formatted: Font:(Default) Times New Roman

Sandhya Tiwari 15/6/14 18:38
Comment [31]: Add the other image for MUSTANG alignment

Sandhya Tiwari 16/6/14 14:16



Deleted: ... [1]

Unknown
Formatted: Font:(Default) Times New Roman

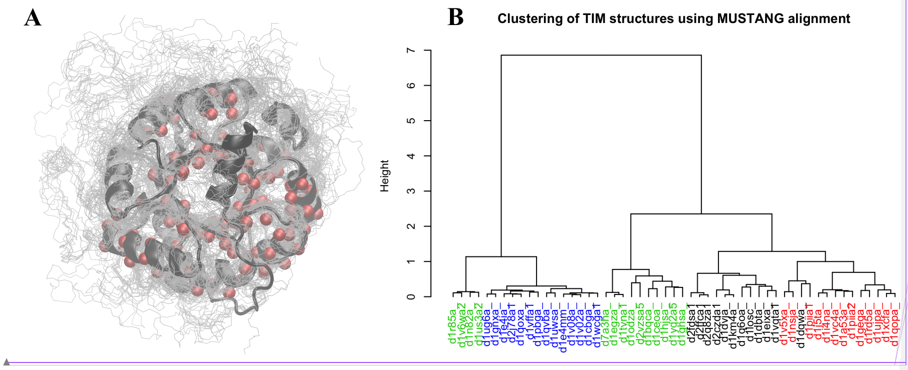


Figure 2. . A) MUSTANG alignment of 53 structures with the TIM-barrel fold. The light grey lines show the superimposition of the structures, while the dark grey cartoon of the secondary structure shows one of them as a representative. The red spheres highlight the points on the structure that are conserved throughout the alignment, which are distributed evenly here. B) K-means clustering (with k = 4) of the Bhattacharya score analysis comparing the covariances using the MUSTANG alignment. 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 good groupings with respect to the family level, but a less distinct separation at the superfamily level with respect to the green group.

A thorough discussion about defining comparable regions of a protein and on some strategies for aligning proteins using models of their intrinsic flexibility is described in a review of Micheletti {Micheletti, 2012 #31}.

4.3. Motion of comparable fragments

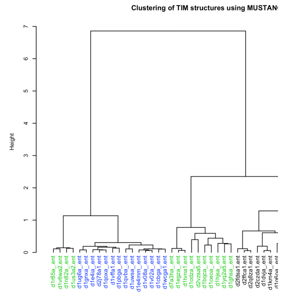
Once it has been defined what parts of the compared proteins correspond to each other, one is typically left with the task of comparing a subset of each structure to each other. Each protein is thus partitioned into a subset of compared atoms A, and a subset of excluded atoms B. While only A has a corresponding part in all the structures compared, B has an influence on the dynamics of A and should be retained in the calculation of motion. 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 deforms 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,

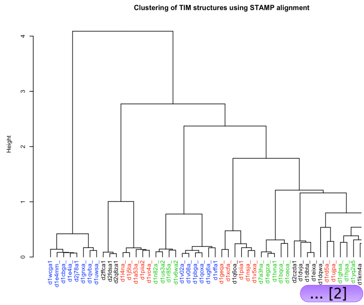
Sandhya Tiwari 16/6/14 14:16
Comment [32]: Will be fixed for readability

Unknown
Formatted: Font:(Default) Times New Roman

Sandhya Tiwari 15/6/14 18:38
Comment [33]: Will be fixed for readability

Sandhya Tiwari 16/6/14 14:16
Deleted:





... [2]

Unknown
Formatted: Font:(Default) Times New Roman

Unknown
Formatted: Font:(Default) Times New Roman

Edvin Fuglebakk 16/6/14 09:32
Comment [34]: I think Micheletti discusses pair-wise alignments, if that is the case it should probably be stressed.

Edvin Fuglebakk 16/6/14 09:40
Comment [35]: I think this can be simplified by moving the formalism to the formalism section. It requires some rewriting, so I suggest we leave it for now, and I will take care of it if you agree it is a good idea.

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:

$$\tilde{\mathbf{H}} = \mathbf{H}_{aa} - \mathbf{H}_{ab}\mathbf{H}_{bb}^{-1}\mathbf{H}_{ab}^{\mathrm{T}} \qquad [8]$$

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.

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 comparable frame of reference. When comparing models of identical proteins, as is done when benchmarking models, a comparable reference frame can be obtained by simply rotating the protein prior to calculation. In the case when different proteins are compared, an approximation to a common reference frame is typically obtained by roto-translating 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 includes considering internal coordinates {Mendez, 2010 #8}, or comparing rotationally invariant properties of the normal modes, like the correlation matrix in Eq. [7].

Sandhya Tiwari 16/6/14 11:22
Formatted: Line spacing: 1.5 lines

Sandhya Tiwari 16/6/14 11:22
Deleted:

5. Strategies and Applications of Comparative Analysis

Comparing multiple structures has always been a natural extension to the study of dynamics. In the case of ENMs, where it is computationally efficient and inexpensive, there have been a variety of studies comparing sets of structures in various ways. A basic example of such analysis was done by Luo and Bruice, where they conducted a visual inspection of the normal modes vectors on six structures in the dihydrofolate reductase kinetic pathway which they related back to kinetic, NMR and theoretical experiments {Luo, 2009 #155}. There are some other more quantitative examples employed by studies using ENMs to characterise properties derived from multiple structures, whether from the same sequence, related by homology or fold.

5.1 Multiple conformations of the same structure

Previously, the overlap analysis was introduced as a way to understand the transformation between two distinct conformational states of an enzyme, allowing for the identification of modes that explain the conformational changes seen {Marques, 1995 #49;Tama, 2001 #37; Reuter, 2003 #42}. The finding that the lowest frequency modes are often the ones that are functionally relevant supports this kind of analysis. Traditional dynamics studies that compare two extreme states e.g. fully open ligand-free conformations vs. fully closed ligand-bound conformations that leads to interpretations that the modes sampled in the transition of these two states are indeed the ones important for conformational state. 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 transitions. Stein and colleagues performed serial overlap analyses of unbound to bound conformations from over 12,000 structures to assess the cost of conformational change upon ligand binding, and the classical models they fit{Stein, 2011 #156}. In general, transitions are a much more complex affair that require more detailed and rigorous calculations to estimate{Whitford, 2008 #90;Togashi, 2010 #100}.

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 Principle Component Analysis (PCA). PCA has the advantage of reducing the dimensionality of large-scale data into basis vectors (or principle components), ordered based on the level of variance it describes. Normal modes are analogous to principle components of a Molecular Dynamics trajectory, in that they describe collective motions of proteins with eigenvectors that follow an ordering based on the size of the amplitude{Skjaerven, 2011 #53}. Here we focus on using PCA not to reduce the dimensionality of dynamics description, but as a means to reduce information from multiple structures. One strategy employing PCA involves calculating principle components for a large collection of structures for a given protein, before running ENM calculations on representative structures {Katebi, 2014 #52}. Yang and colleagues have shown that 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}.

Sandhya Tiwari 16/6/14 18:45
Deleted: intrinsic motions

Using the ENMs themselves to produce covariance matrices that can be compared between them to get an idea of the difference in dynamics for different states{Seckler, 2013 #83}. ENMs are useful in studying allosteric effects, as it is seen to be able to estimate entropic effects to a degree{Motlagh, 2014 #77}. a

5.3 Co-evolution studies

Lukman and Grant (2009) surveyed maltose transporters for conserved structural motifs, before testing their influence on the overall dynamics of these proteins in different states. <Introduce perturbation analysis on many structures here>

5.4 Multiple subunits of oligomers

In addition, when comparing oligomeric structures, it is important to note the effect of calculating the modes in different states, and making a one-to-one comparison. The ordering of the modes is different between different states, where subunit-subunit motion make up the lowest frequency modes, and thus, we lose information when transforming eigenvectors into comparable measures. Examples of work done to understand oligomeric state, comparatively: amino acid kinase family{Marcos, 2011 #116}, GPCRs{Niv, 2008 #115}, serine receptors Tsr{Hall, 2012 #111}, p53{Kantarci, 2006 #109}.

5.5 Homologues and structural fold

Structure comparison has been long established as a means to understanding the evolution of proteins. The link between sequence and structure has been long studied since the work of Lesk and Chothia {Chothia, 1986 #110}. In many cases, even from visual inspection, a structure is seen to not just encode dynamic information, but a historical time-point in the evolution of a given family of proteins. Sequence mutations, insertions and deletions are accommodated by the plastic deformations of a common architectural core, and that this core, or fold, is able to retain the precise geometry of the active site, even if peripheral regions or accessory domains vary{Hasegawa, 2009 #118}.

Since the collective motions of ENMs are basically a function of the protein shape, these models are ideal tools for studying the role of protein fold in functional motion. Apart from isolating the structural properties from the chemical details stabilizing the fold, the ENMs are also computationally inexpensive, and can be applied to large data sets of structures, if models with transferrable parameters are used. Since proper

Sandhya Tiwari 16/6/14 18:51
Moved (insertion) [1]

Edvin Fuglebakk 12/6/14 22:55
Comment [36]: ref: Berezovsky and Mitternacht and scissor model people

Sandhya Tiwari 16/6/14 18:52
Deleted: .. ... [3]

Sandhya Tiwari 16/6/14 19:26
Deleted: In this setting the decomposition of motion provided by elastic network models provides a convenient formalism for presenting general mechanistic models.

Sandhya Tiwari 16/6/14 18:51
Moved up [1]: Using the ENMs themselves to produce covariance matrices that can be compared between them to get an idea of the difference in dynamics for different states{Seckler, 2013 #83}.

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 since proteins are flexible entities elucidating their principal modes of motion is an important ingredient in understanding functional properties of the fold. … In fact the observation that low energy normal modes so frequently appear in functional motion {Nicolay, 2006 #16} has motivated investigations into their evolutionary conservation {Maguid, 2005 #17;Maguid, 2006 #18;Maguid, 2008 #19} and their similarity with the space explored by evolution {Leo-Macias, 2005 #20;Leo-Macias, 2005 #21;Velazquez-Muriel, 2009 #22;Raimondi, 2010 #23}. Interestingly, even if these modes have been shown to be conserved, it has been difficult to show that they are conserved due to a selective pressure apart from any selection pressure on the structure in general {Echave, 2008 #24}. Rather the low energy normal modes are more robust to random structural perturbations and this can be seen as the reason for their conservation {Echave, 2010 #25;Echave, 2012 #26}. An important aspect to consider in this regard is that the normal modes mainly describe the principal directions of collective motion, and the steric constraints on deformations. The energy of conformational transitions can be modulated in many other ways, even when the direction of the transition is in good agreement with ENM predictions. Other techniques could potentially be used to investigate selective pressure on collective degrees of freedom {Fang, 2014 #27}.

Edvin Fuglebakk 1/6/14 22:05
Comment [37]: Get some examples of papers explaining function in terms of normal modes.

Edvin Fuglebakk 1/6/14 22:05
Comment [38]: more refs

Zheng and Doniach{Zheng, 2003 #96} studied the conservation of dynamics in motor proteins by comparing the conformational changes experienced by myosin, F1-F0 ATPase and kinesin. In this case, the authors concluded that the while the large conformational change seen in myosin and F1-F0 ATPase was consistent with a power-stroke type movement, they found that the kinesin followed a Brownian ratchet-type mechanism. Studies like this have also prompted initiatives to categorise protein structures dynamically {Warren, 2014 #89}. Large-scale dynamics database initiatives, such as ProMode, MolmovDB have also shown space for ENMs to characterise structural dynamics that can complement existing structural classifications systems.

…

Edvin Fuglebakk 14/6/14 01:30
Comment [39]: Comparisons in coarser grained models? Like the scissor model, with only a handful of parameters.

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 is also available {Hollup, 2005 #472;Suhre, 2004 #971;Eyal, 2006 #282;Seo, 2012 #907;Kruger, 2012 #571;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}.

…

7. Conclusion/Perspectives

The direction that ENMs and comparing them should head towards.

Sandhya Tiwari 16/6/14 18:18
Deleted: ,

Sandhya Tiwari 16/6/14 18:18
Deleted:

Edvin Fuglebakk 14/6/14 01:31
Comment [40]: Mention which implement comparative part. Loos, prody? Bio3D ?, webnma. Need to make sure I am exhautive if I add this.

Figure captions

Figure 1: STAMP and MUSTANG alignments 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.

Figure 2: Influence of the structure alignment on similarity measures. K-means clustering (with k = 4) of the Bhattacharya score analysis comparing the covariances using A) MUSTANG and B) STAMP alignments. The structures with the TIM Barrel fold are from 2 different SCOP superfamilies, with structures from 2 families from each superfamilies. 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. A) We see good clustering at the family level except for green, while most structures separate at the superfamily level except for most of the green structures. B) We see a heterogenous clustering of the structures.

Highlights