Using a Database of Alternative Conformations for the Large-Scale Analysis of Sequence Features in Flexible Protein Elements: An Exploratory Study

Given a large number of proteins, is there a way to systematically and automatically identify flexible protein elements?

→ 'flexibility' includes more than just hinges and disordered regions

Relative to more rigid elements, what sequence features charachterize the most flexible elements?

→ conservation (inter-species)
→ conservation (intra-species)
→ enrichment or depletion of disease-associated variants
→ enrichment or depletion of cancer-associated variants
→ sequence complexity/entropy

Background

Single Nucleotide Polymorphisms & Structure

Amino acid substitutions are frequently introduced to proteins artificially, as a means of probing the consequences of structural perturbations on allostery, kinetics, secondary structural elements, stability, etc.

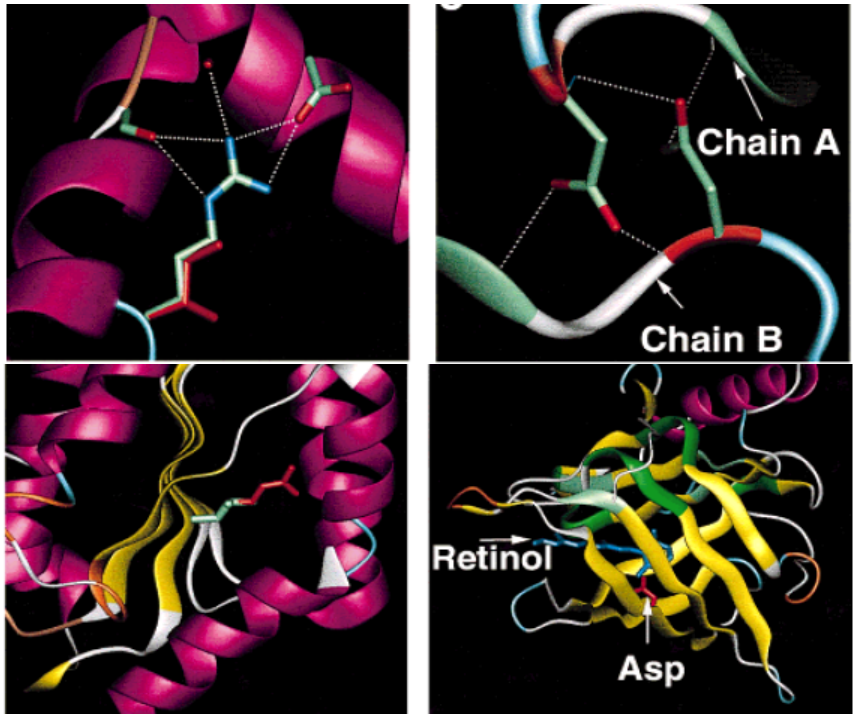
Missense SNPs & disease as a consequence of impaired functionality/ regulation

1) coagulation factor XIIIA: disrupts salt bridge

2) aldehyde dehydrogenase: burying charged residue in h-phobic core

3) aldolase: destroy H-bonds needed as part of quartenary structure

4) Vitamin A & retinol binding protein: interferes w/ligand binding



Wang,Z. and Moult,J. (2001) SNPs, protein structure and disease. Hum. Mutat., 17, 263–270.

Background

Previous work on flexibility analysis in the Gerstein lab

Structure(s) submitted by user

↓

Alignment & Superposition

↓

Orientation

↓

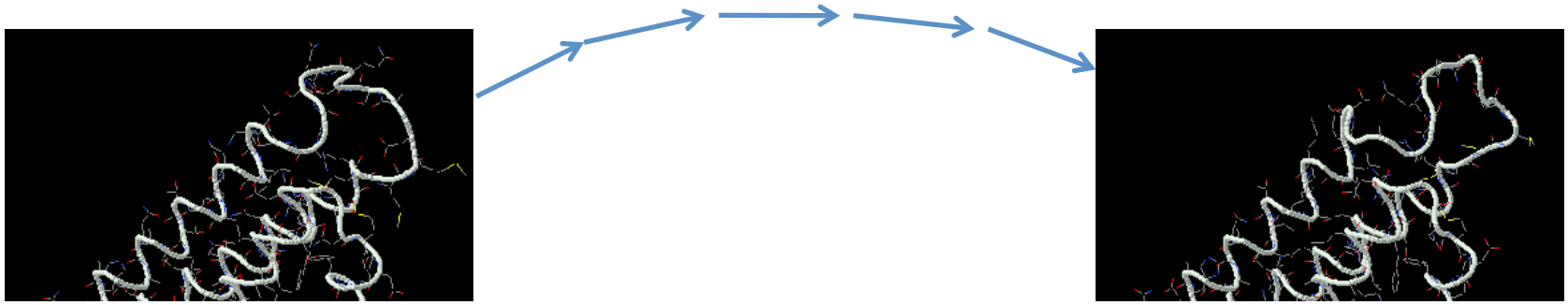
Homogenization of Terminal Structures

↓

Interpolation

↓

Viewing & Web Report



Grow the repertoire of morphs in MolMovDB to provide more data, and more comprehensively represent nature's pool of protein movements

→ Largely already finished: may use distinct conformational states identified in our group (MMDB, STRESS)

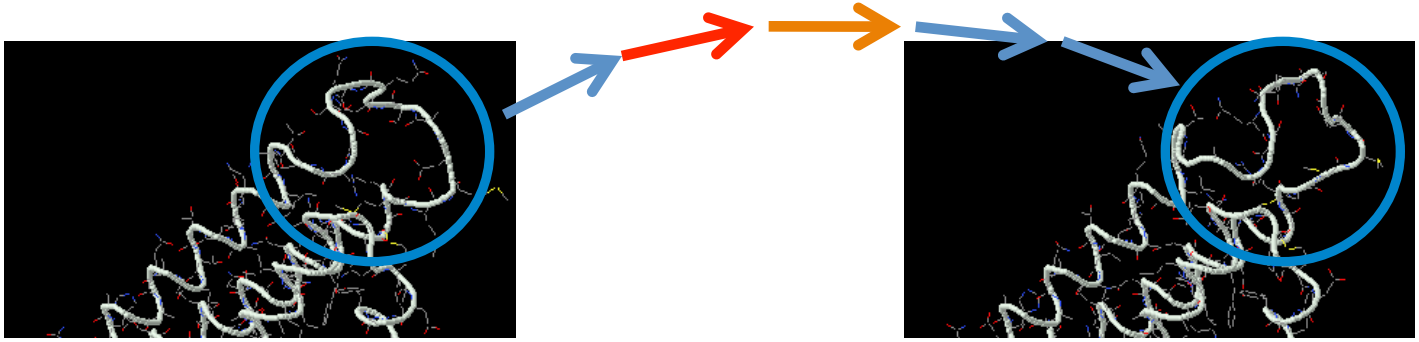
Devise means of identifying flexible elements

→ "conventionally-defined" all-atom dihedral angles in the polypeptide backbone

→ Cα – Cα dihedral angles

→ Use 'terminal structures' (ie, X-ray crystal structures), although this is not a perfect representation (some flexible elements will remain hidden).

N	0.3	0.5	-1.4	-1.6	-2.0	-0.8	-0.5	…..	0.3	C
N	0.2	0.4	-0.8	-1.9	-1.7	-2.1	0.4	…..	0.3	C



Potential Experimental Plans

1) For one protein, use Python scripting to calculate dihedral angle changes in the protein backbone of one single-chain protein for which 2 PDB structures are available. Maybe calculate "conventional" all-atom backbone dihedral angle changes or the changes defined using Cα – Cα "bonds". Confirm flexibility 'identifications' by visualizing morphs or aligned structures.
2) For that protein, devise means of evaluating cross-species conservation at the level of each residue (ConSurf has previously been used in this lab – use ConSurf or some alternative means of quantifying conservation).
3) Scale the above analysis to a large number of proteins. Perhaps use existing alternative conformers in MolMovDB, or alternatively those from STRESS.
4) Evaluate relative conservation of highly flexible vs highly 'rigid' elements using intra-species datasets for human sequence data (1000 Genomes, ExAC) [SK]
5) Disease association: search for enrichment/depletion of HGMD SNPs or SNVs in cancer datasets [SK]

A Few Caveats & Considerations

- The “motions” are interpolated, not empirical. Is there a bias against ‘flexibility’ built into the interpolation algorithms, for instance?

- Bias in X-ray structures excludes many disordered elements

- Different classes of flexible elements (much of flexibility is directly functional, some of it is not) – how to deal w/biological heterogeneity?

- Are (local) motions biologically accurate, or do they result from crystallization, poor resolution, or other artifacts – how to discriminate?

- Establishing thresholds