More from your Multiple Sequence Alignment: *In silico* second site suppression?

Greg Gloor

September, 2003

Current co-conspirators

Stan Dunn

Lindi Wahl

Louise Martin

Kaizhong Zhang

Jana Moulin

Colleen Ball

Big picture

Trying to identify and display important amino acid residues in protein families in-silico

Today's snapshot

Can identify functionally important residues, even if not conserved and display these on structure

Using mutual information (co-evolution, co-variation, in-silico second-site suppression, in-silico two-hybrid analysis)

Mutual information or co-variation studies

- Chiu (1991) - MI of NA sequences for structure prediction
- Korber et al (1993) - MI of the HIV V3 loop
- Clarke (1995) - MI of homeodomains
- Atchley et al (2000) - MI of bHLH
- Valencia (2002) - MI as in silico two-hybrid screen

Methods to identify important residues

- Mutagenesis (Genetics/Molecular Biology)
- Chemical alteration (Biochemistry)
- Multiple Sequence alignment (Bioinformatics)
- Structure Determination (Biochemistry)

Mutagenesis / Alteration

- Labor intensive
- Only hit some sites
- Results depend upon the method used

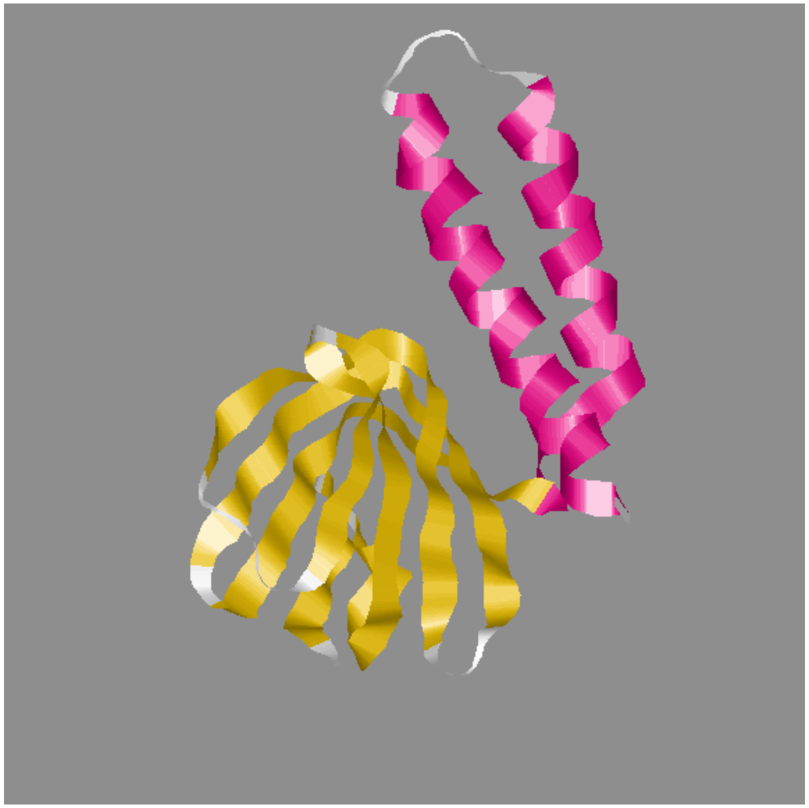
Mutagenesis / Alteration

- Labor intensive
- Only hit some sites
- Results depend upon the method used

- But … Results provide functional detail

Protein structure

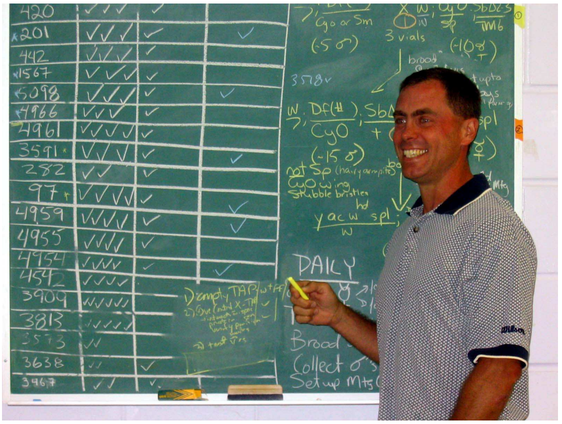
- Gold standard for fine detail examination of protein structure
- Feeling is sometimes that the structure is an end point
- But is only a snapshot



Protein structure

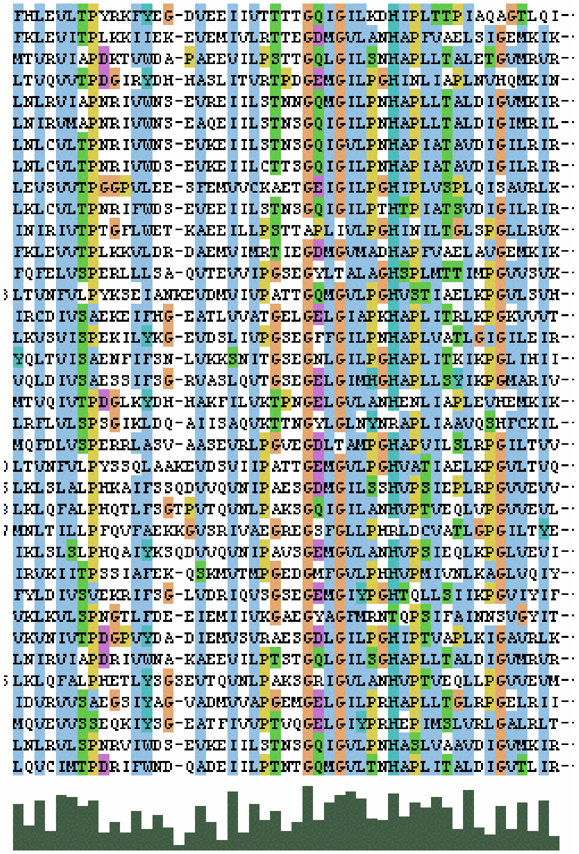
Protein structure

Protein structure



Bioinformatic: Multiple sequence alignment

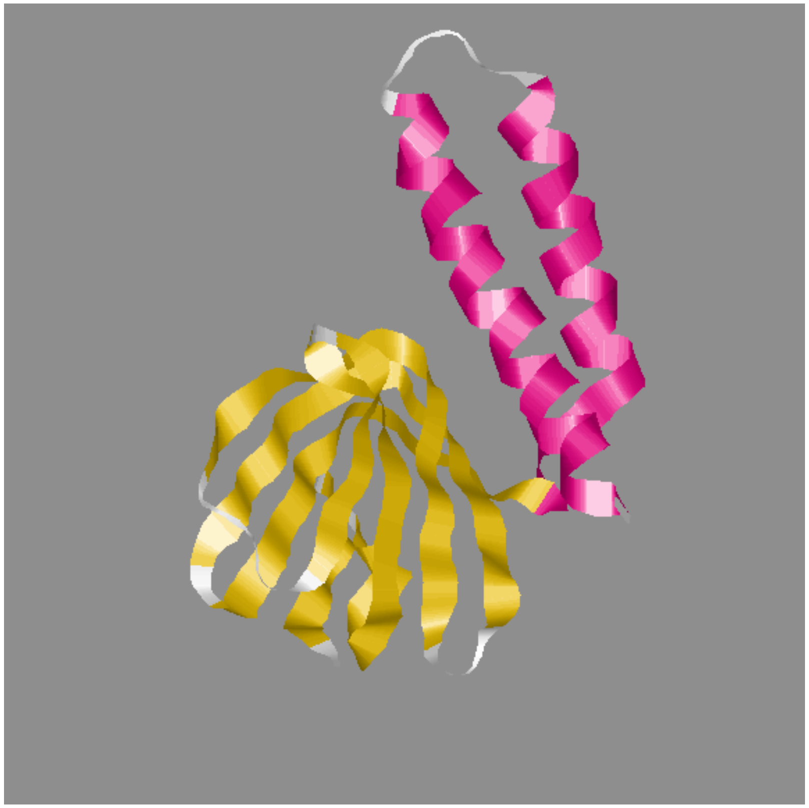
- Treats each position independently
- Conserved residue positions - lots of info
- Gapped positions - little info
- Variable positions - often hard to pick out a pattern



```
FHLEVLTPYRKFYEG-DVEEIIVTTTTGQIGILKDHIPLTTPIAQAGTLQI--
FKLEVITPLKKIIEK-EVEMIVLRTTEGDMGVLANHAPFVAELSIGEMKIK--
MTVRVIAPDKTVWDA-PAEEVILPSTTGQLGILSNHAPLLTALETGVMRVR--
LTVQVVTPDGIRYDH-HASLITVRTPDGEMGILPGHINLIAPLNVHQMKIN--
LNLRVIAPNRIVWNS-EVREIILSTNNGQMGILPNHAPLLTALDIGVMKIR--
LNIRVMAPNRIVWNS-EAQEIILSTNSGQIGILPNHAPLLTALDIGIMRIL--
LNLCVLTPNRIVWNS-EVKEIILSTNSGQIGVLPNHAPIATAVDIGILRIR--
LNLCVLTPNRIVWDS-EVKEIILCTTSGQIGILPNHAPIATAVDIGILRIR--
LEVSVVTPGGPVLEE-SFEMVVCKAETGEIGILPGHIPLVSPLQISAVRLK--
LKLCVLTPNRIFWDS-EVEEIILSTNSGQIGILPTHTPIATSVDIGILRIR--
INIRIVTPTGFLWET-KAEEILLPSTTAPLIVLPGHINILTGLSPGLLRVK--
FKLEVVTPLKKVLDR-DAEMVIMRTIEGDMGVMADHAPFVAELAVGEMKIK--
FQFELVSPERLLLSA-QVTEVVIPGSEGYLTALAGHSPLMTTIMPGVVSVK--
LTVNFVLPYKSEIANKEVDMVIVPATTGQMGVLPGHVSTIAELKPGVLSVH--
IRCDIVSAEKEIFHG-EATLVVATGELGELGIAPKHAPLITRLKPGKVVVT--
LKVSVISPEKILYKG-EVDSLIVPGSEGFFGILPNHAPLVATLGIGILEIR--
YQLTVISAENFIFSN-LVKKSNITGSEGNLGILPGHAPLITKIKPGLIHII--
VQLDIVSAESSIFSG-RVASLQVTGSEGELGIMHGHAPLLSYIKPGMARIV--
MTVQIVTPDGLKYDH-HAKFILVKTPNGELGVLANHENLIAPLEVHEMKIK--
LRFLVLSPSGIKLDQ-AIISAQVKTTNGYLGLNYNRAPLIAAVQSHFCKIL--
MQFDLVSPERRLASV-AASEVRLPGVEGDLTAMPGHAPVILSLRPGILTVV--
LTVNFVLPYSSQLAAKEVDSVIIPATTGEMGVLPGHVATIAELKPGVLTVQ--
LKLSLALPHKAIFSSQDVVQVNIPAESGDMGILSSHVPSIEPLRPGVVEVV--
LKLQFALPHQTLFSGTPVTQVNLPAKSGQIGILANHVPTVEQLVPGVVEVL--
MNLTILLPFQVFAEKKGVSRIVAEGREGSFGLLPHRLDCVATLGPGILTYE--
IKLSLSLPHQAIYKSQDVVQVNIPAVSGEMGVLANHVPSIEQLKPGLVEVI--
IRVKIITPSSIAFEK-QSKMVTMPGEDGMFGVLPHHVPMIVNLKAGLVQIY--
FYLDIVSVEKRIFSG-LVDRIQVSGSEGEMGIYPGHTQLLSIIKPGVIYIF--
VKLKVLSPNGTLFDE-EIEMIIVKGAEGYAGFMRNTQPSIFAINNSVGYIT--
VKVNIVTPDGPVYDA-DIEMVSVRAESGDLGILPGHIPTVAPLKIGAVRLK--
LNIRVIAPDRIVWNA-KAEEVILPTSTGQLGILSGHAPLLTALDIGVMRVR--
LKLQFALPHETLYSGSEVTQVNLPAKSGRIGVLANHVPTVEQLLPGVVEVM--
IDVRVVSAEGSIYAG-VADMVVAPGEMGELGILPRHAPLLTGLRPGELRII--
MQVEVVSSEQKIYSG-EATFIVVPTVQGELGIYPRHEPIMSLVRLGALRLT--
LNLRVLSPNRVIWDS-EVKEIILSTNSGQIGVLPNHASLVAAVDIGVMKIR--
LQVCIMTPDRIFWND-QADEIILPTNTGQMGVLTNHAPLITALDIGVTLIR--
```

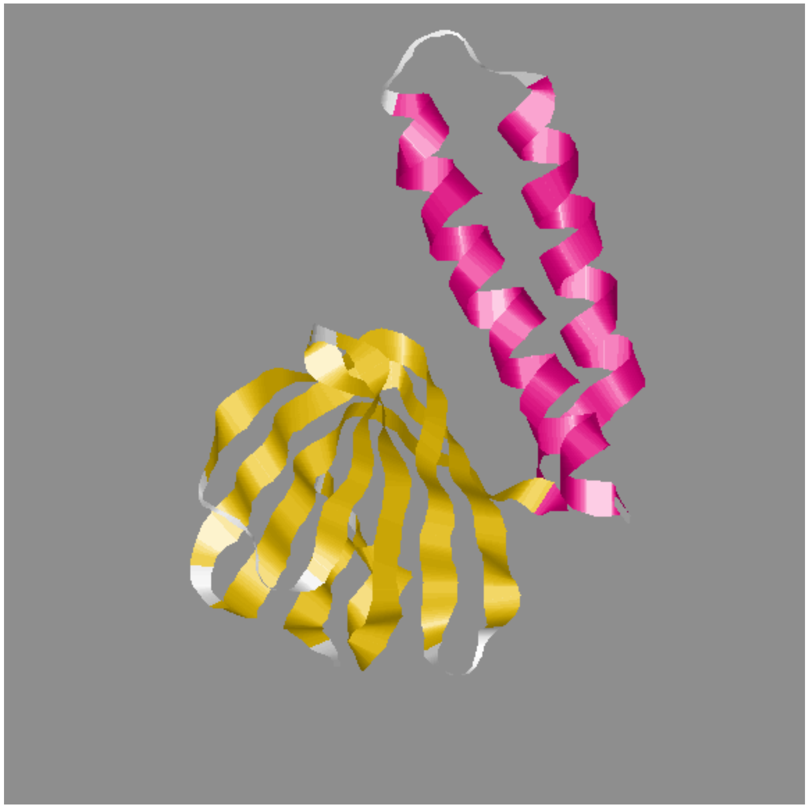
Protein structure

- How can we map the functional and the bioinformatic information onto this structure?
- Is there other information to be mapped?



Protein structure

- Really want to get at underlying relationship between sequence and structure
- Residues interact with each other within and between proteins
- Which residues interact?
- How important are these interactions?



Another Layer: Mutual Information

- Entropy: measures the amount of disorder of a single column or at a pair of columns.
- Maximum when all 20 aa at equal frequencies in a column, minimum when the column is completely conserved

- Mutual Information: measures the uncertainty of a second position given that we know the first position.
- Maximum when there is only one aa2 partner for each aa1
- Minimum when each aa1 has all possible partners at position 2

Work flow to find MI

1. Determine all sequence homologs , **PSI-BLAST**
2. Gather sequences for MSA from BLAST output, **PSI_parse.pl**
3. Produce MSA , **MAFFT, CLUSTAL** or T-Coffee
4. Remove identical sequences for sequence sets, **Jalview**
5. Generate co-variance matrix and Z score , **covary.pl, peerZ.pl** (sum, pair, product)
    1. generates the raw MI file
    2. calculates the peer Z score for the MI
6. Calculate random co-variance (if desired), **entropy_vs_covary.pl**
    1. calculates the average MI due to random chance for the MSA given the aa frequencies at each position. This can then be fed into the covary.pl and peerZlpl programs. (This is Lindi's idea! My implementation is in Perl so is slow.)
7. Make batch files to color residues by max MI score, **color_rasmol.pl**, and conserved residue positions, **color_ent.pl**
8. Visualize results by channeling output to **RASMOL**

Model system: ATP synthase epsilon subunit

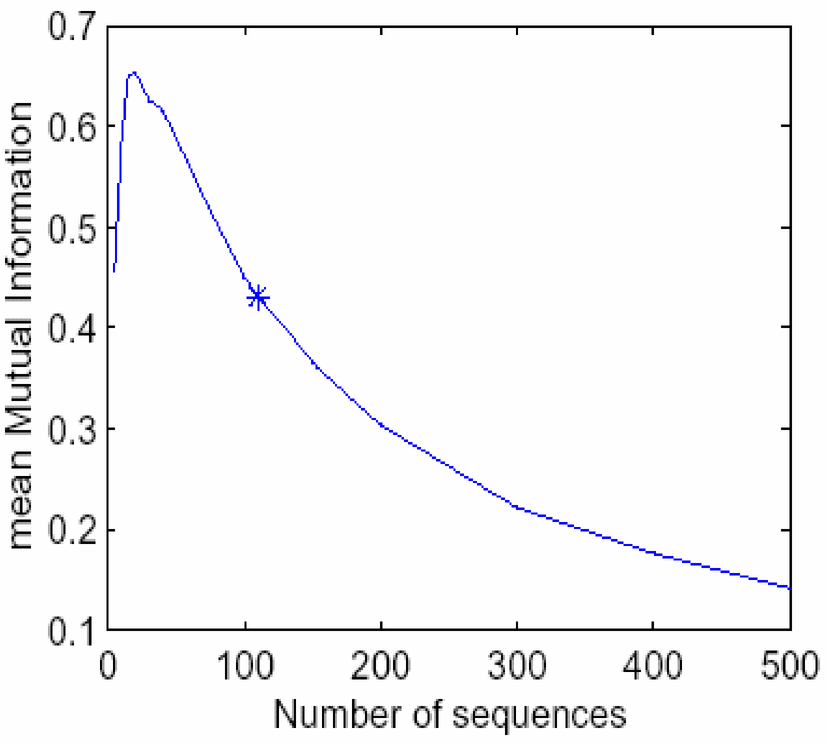
- Ubiquitous multi-protein complex
- Converts proton gradient into mechanical motion, then into ATP
- Epsilon subunit (gray) attaches gamma (blue) to membrane subunit (a)
- Variable enough to show co-evolution
- Conserved between kingdoms

QuickTime™ and a Video decompressor are needed to see this picture.

A side view of alpha3-beta3-gamma
By Hongyun Wang & George Oster, U.C.Berkeley

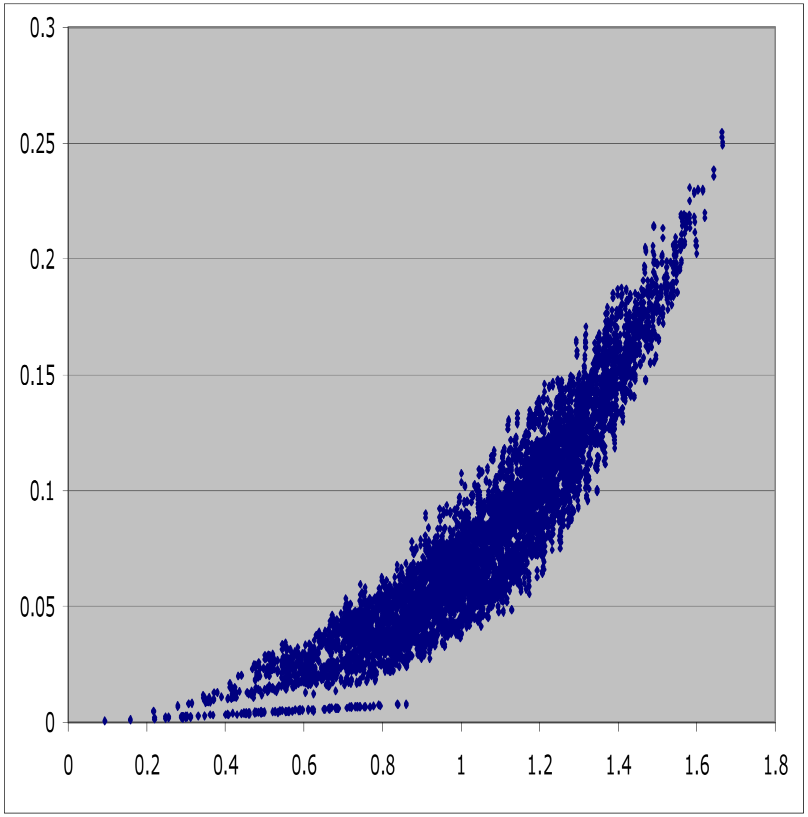
Sources of MI

- Chance
- Shared ancestry
- Structure-function linkage



MI of epsilon aa positions by random chance

- Each dot represents the MI of a pair of aa residues
- Plot is calculated random MI vs sum of entropy of the two residues
- No shared ancestry
- No structure-function linkage
- Plot is rather symmetrical at any given entropy sum

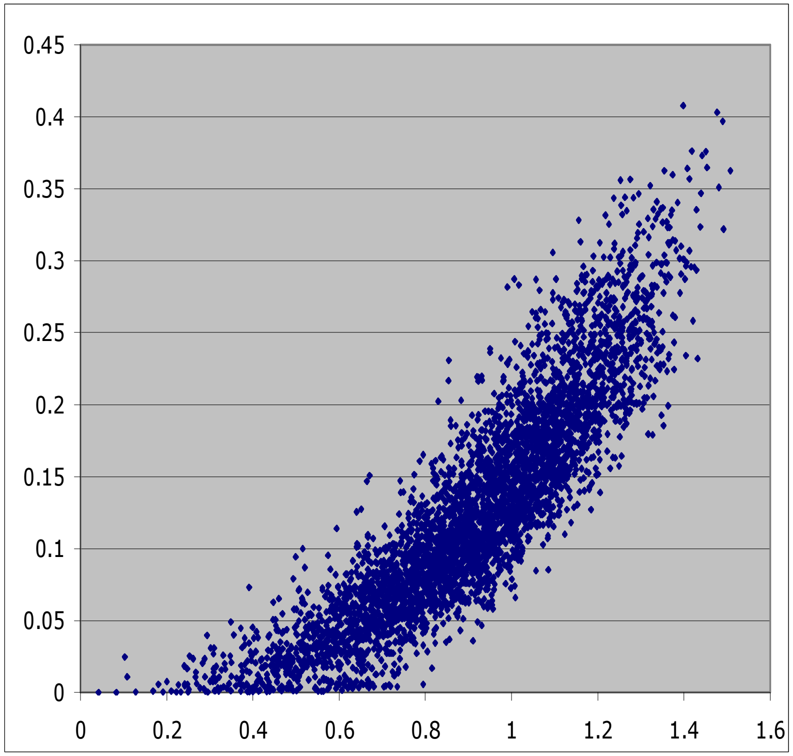


Real MI of epsilon aa positions

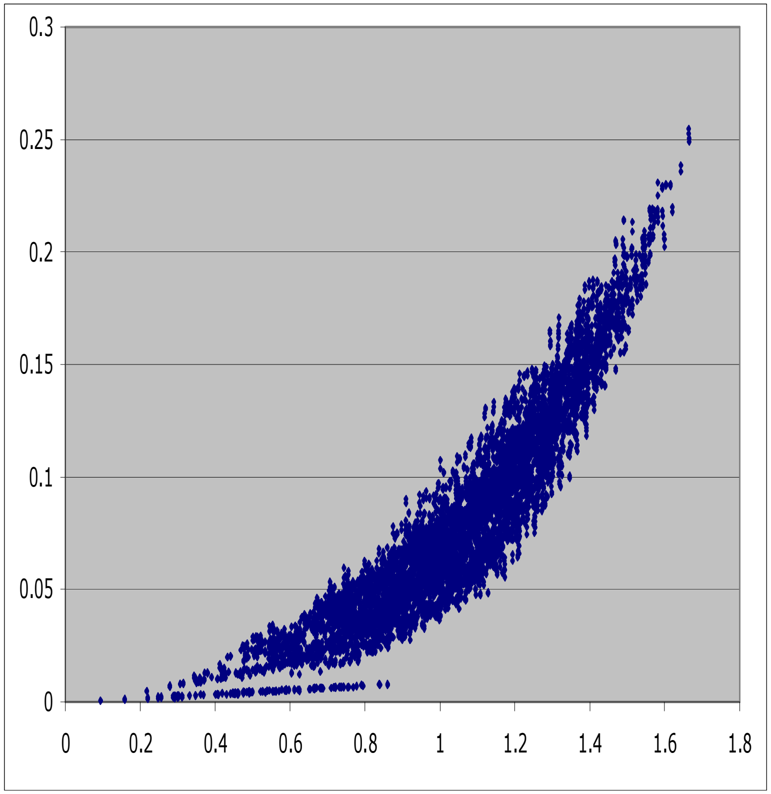
- MI vs sum of entropy at two positions

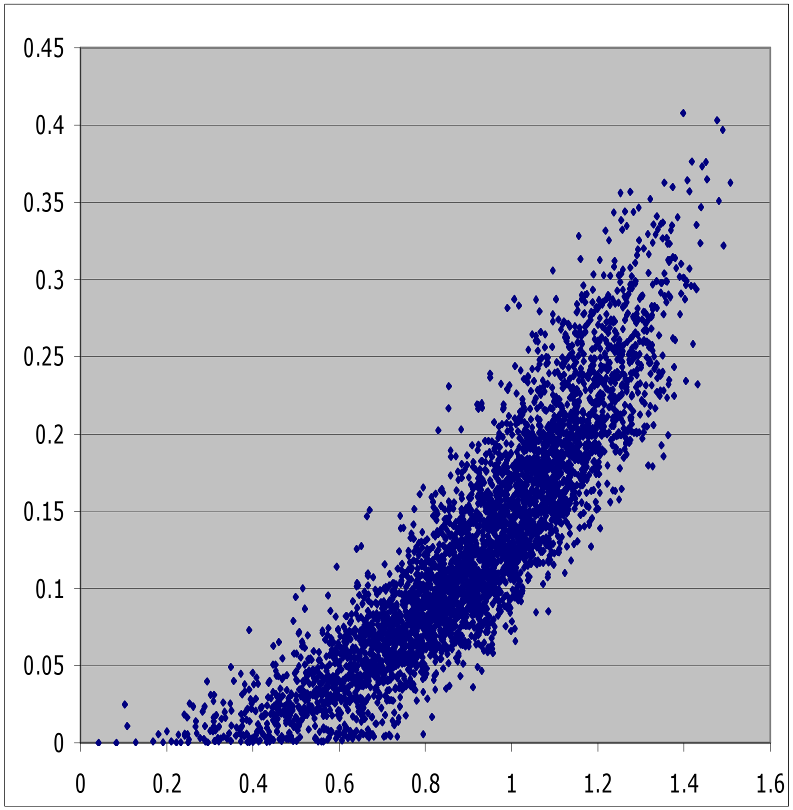
- Shared ancestry
- Structure-function linkage

- Plot is less symmetrical at any given entropy sum,
- Certain pairs of residues stand above the crowd
- Fewer pairs of residues are below the crowd



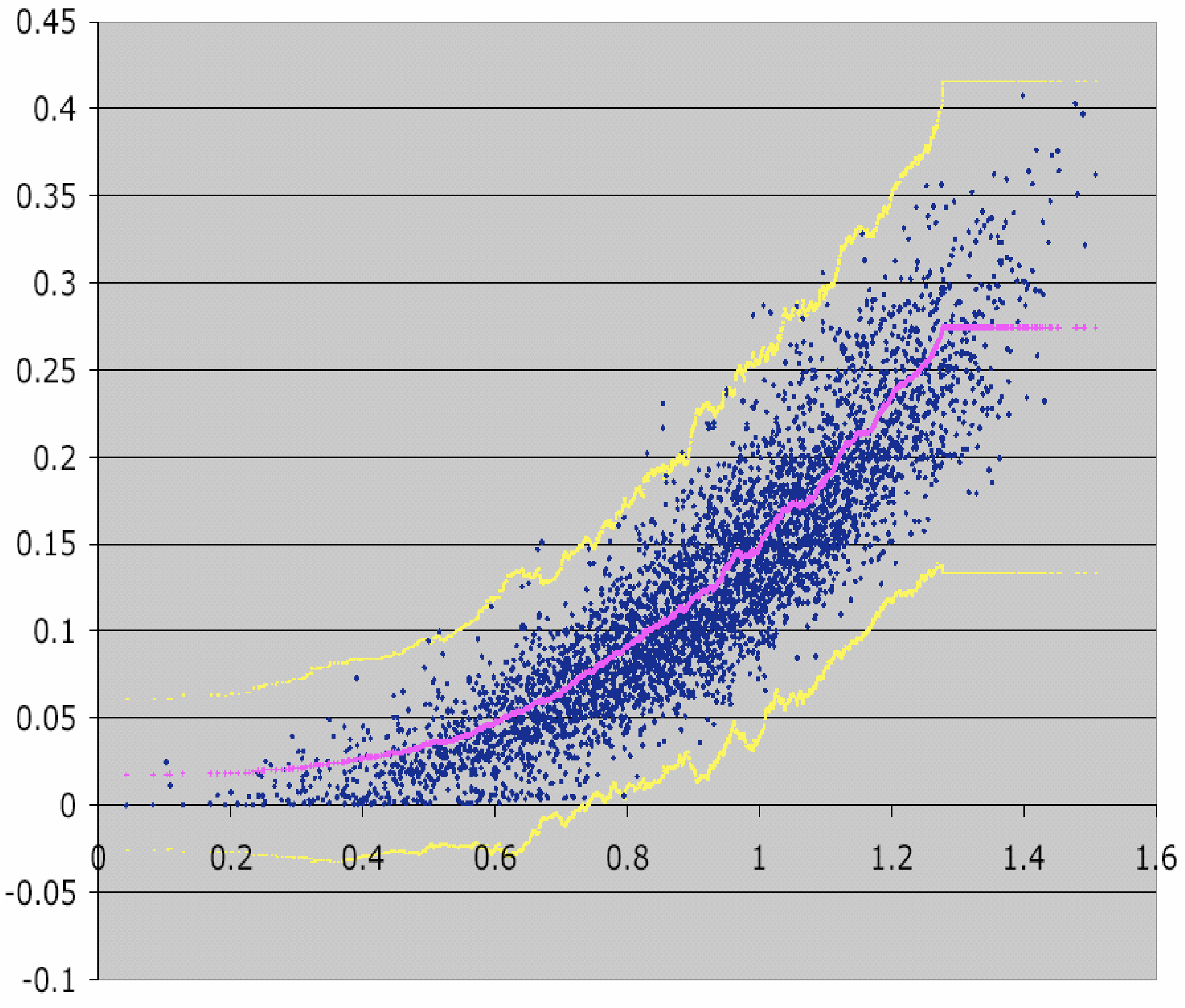
Random and real MI of epsilon aa positions





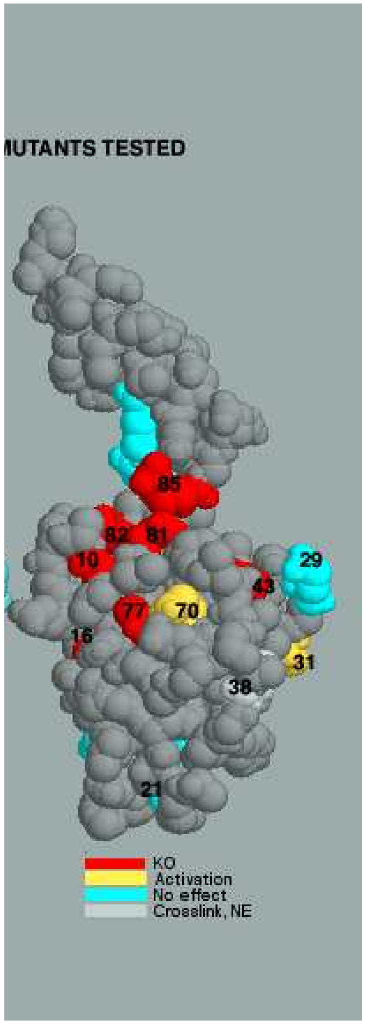
Data manipulation (288 sequences in MSA)

- Find entropy peers (+/- 100)
- Calculate mean, SD for entropy window
- Calculate Z score
- Plot Z scores on structure
- 14 residue pairs at or above Z = 3



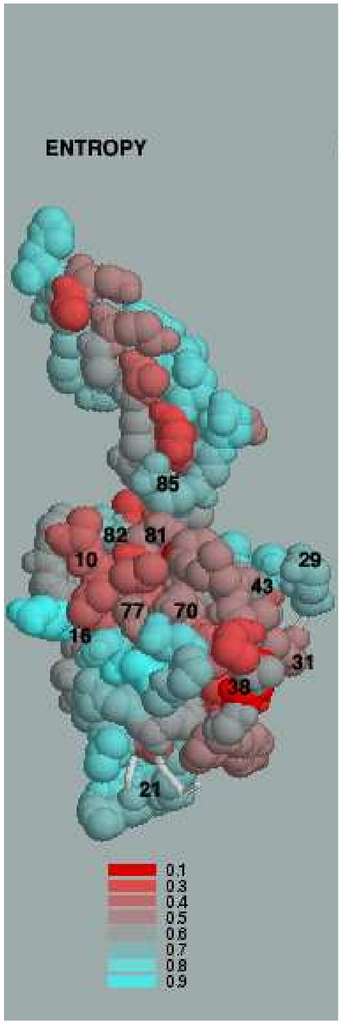
Residues tested in epsilon-gamma interactions (front)

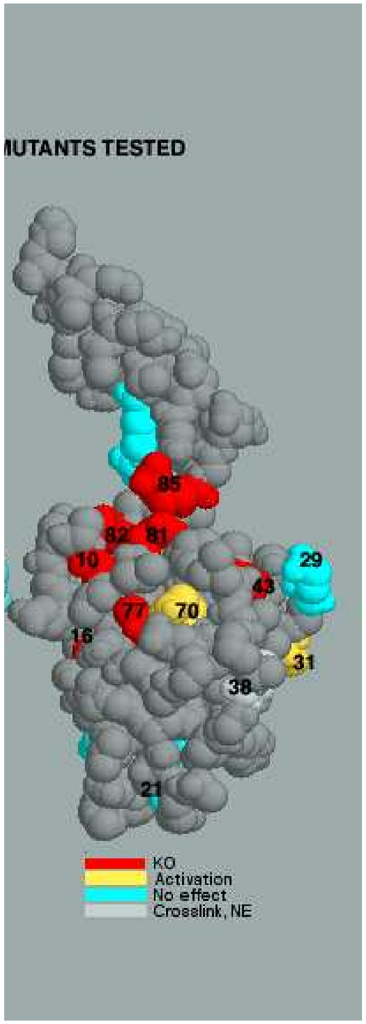
- Vik et al has mapped some positions important for E activity (orange up, red down, cyan no effect, white X-link)



Residue entropy vs functional importance(front)

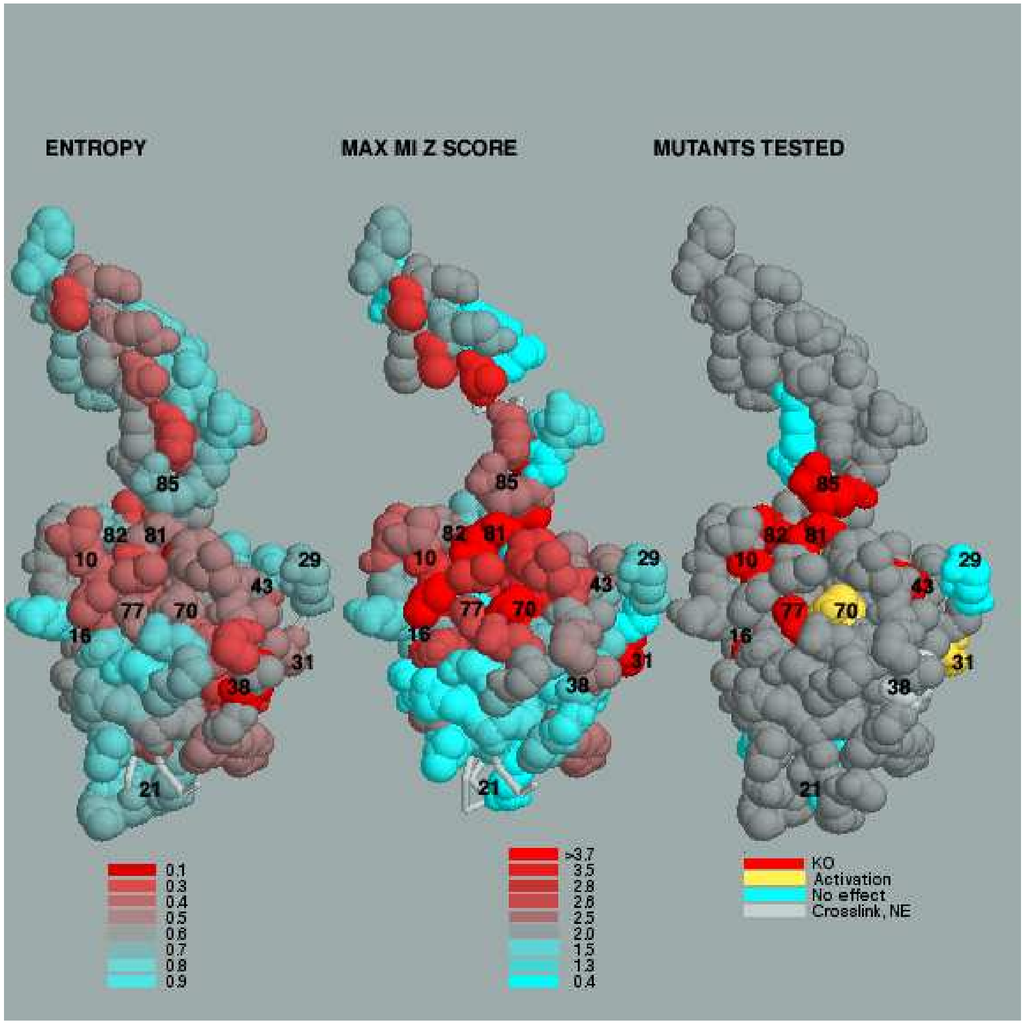
- Residues with functional importance are often not conserved (i.e. moderate to high entropy



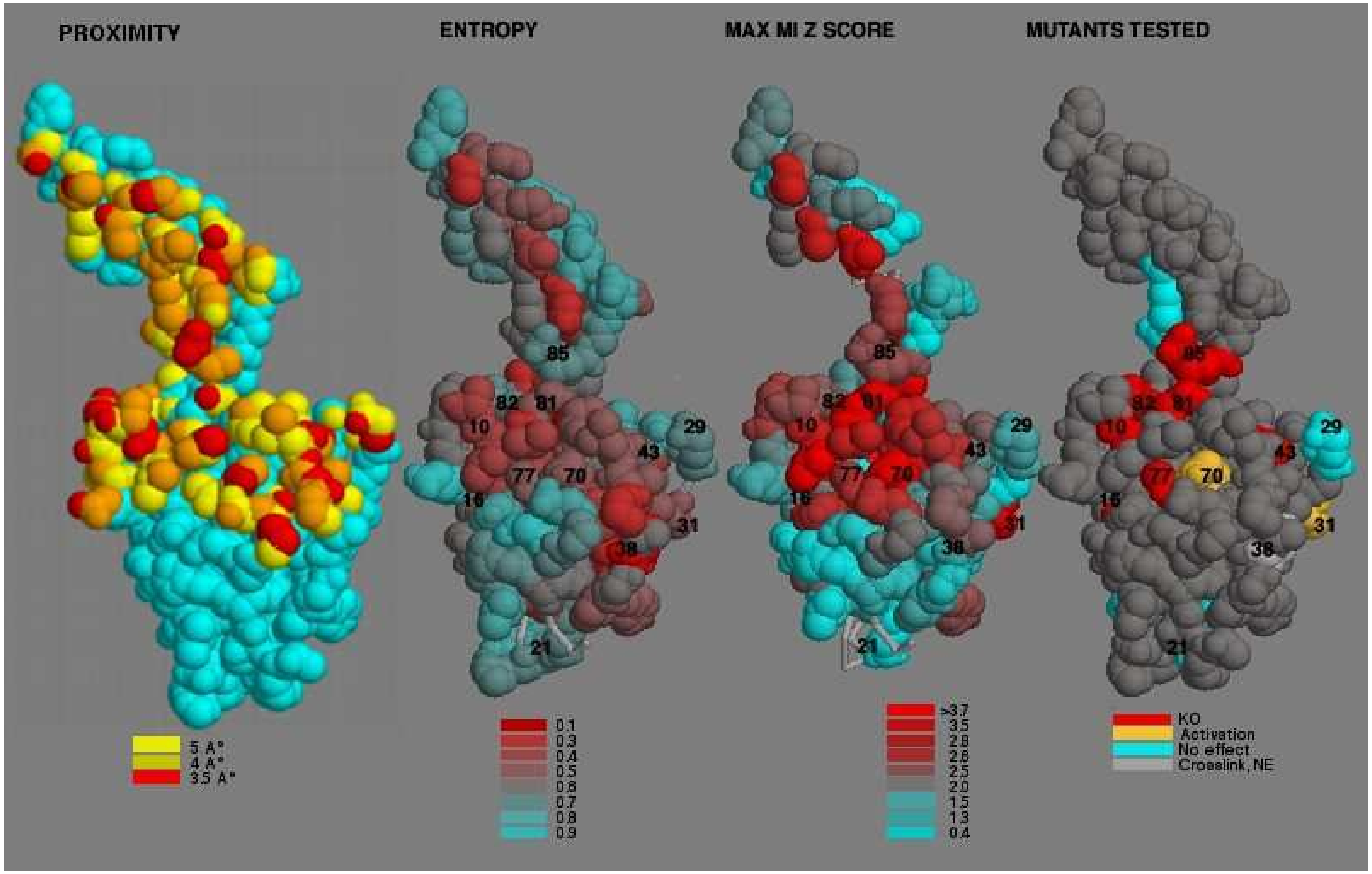


Entropy vs MI vs Function (front)

- MI able to predict mutagenesis outcome in most cases



Proximity vs entropy vs MI vs mutants

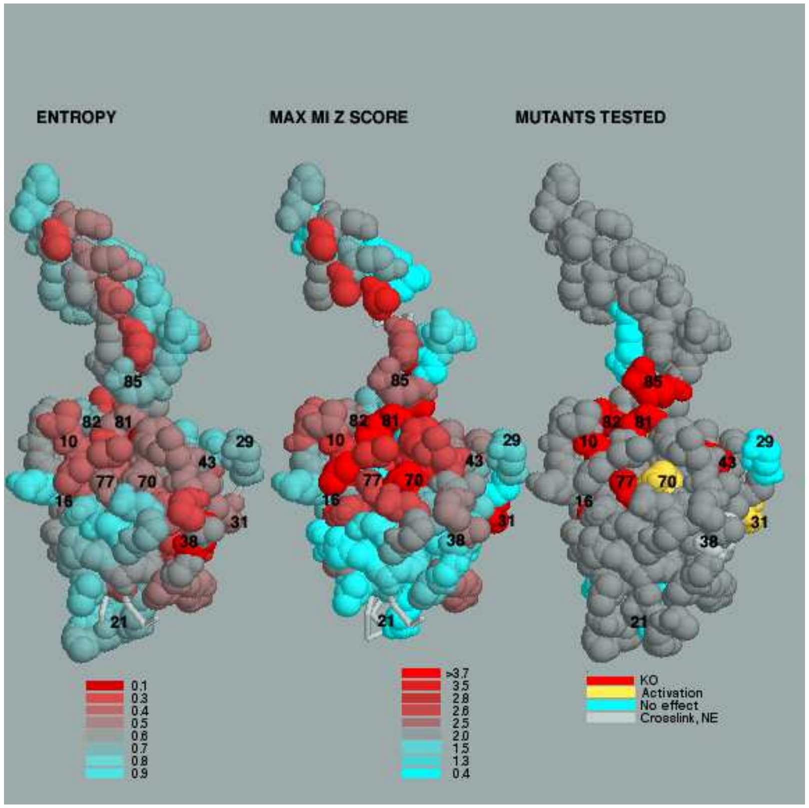


Summary

- *MI requires lots of sequences to give information*
- *MI is a poor method to find pairwise residue interactions*
- *MI is widespread in proteins*
- *MI gives evidence about major function(s) of the protein*
- *We are able to identify and categorize almost all the residues in epsilon that were previously mutagenized*

Information

- Have
  - Entropy (conservation)
  - MI (max)
  - Mutant effect
  - Structure
  - Proximity
- Don’t Have
  - Possible residues (MSA, & mutagenesis)
  - Binding sites
  - Other information



Open Problems

- *Which MI values can we trust?*
  - Significance of the results is an ongoing problem
- *What is the grouping limit?*
  - How do we put interacting residues into networks
- *What does this tell us about proteins*?
  - Evolution, function, structure
- *How to put all of the information together*
  - Three views (at least) is clearly not the best way
  - what about proteins without structures?