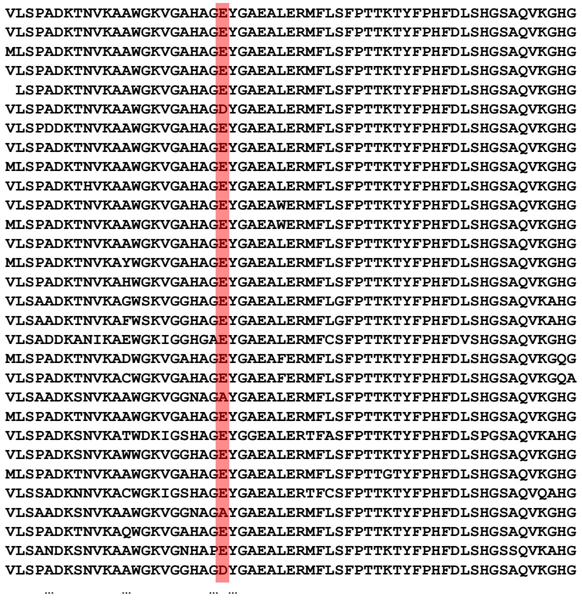
Correlated Mutations – structure prediction

Normal lectures

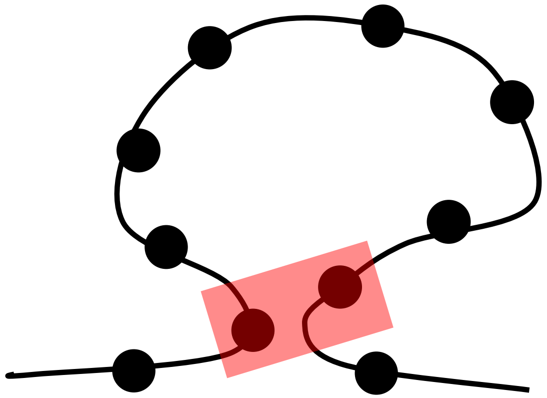
- multiple sequence alignments – 99 % of our analysis
    - columns are independent of each other

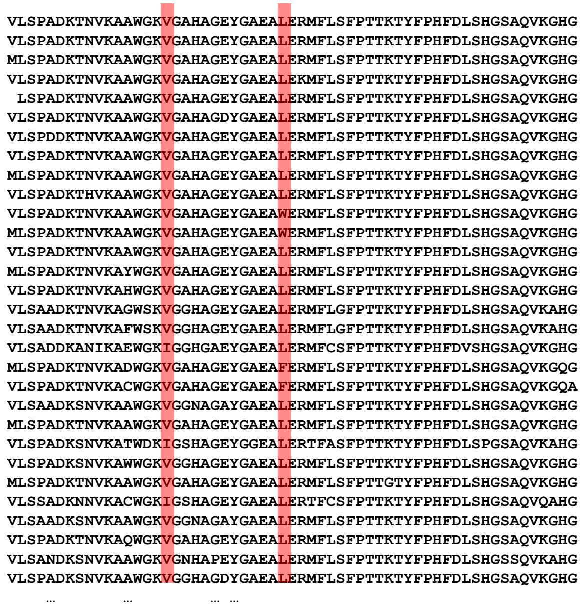


```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALEKMFLSFPTTKTYFPHFDLSHGSAQVKGHG
 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTHVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAYWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAHWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSAADKTNVKAGWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSAADKTNVKAFWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSADDKANIKAEWGKIGGHGAEYGAEALERMFCSFPTTKTYFPHFDVSHGSAQVKGHG
MLSPADKTNVKADWGKVGAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQG
VLSPADKTNVKACWGKVGAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQA
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKSNVKATWDKIGSHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHG
VLSPADKSNVKAWWGKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTGTYFPHFDLSHGSAQVKGHG
VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAHG
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAQWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSANDKSNVKAAWGKVGNHAPEYGAEALERMFLSFPTTKTYFPHFDLSHGSSQVKAHG
VLSPADKSNVKAAWGKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
    …       …        … …
```

Here

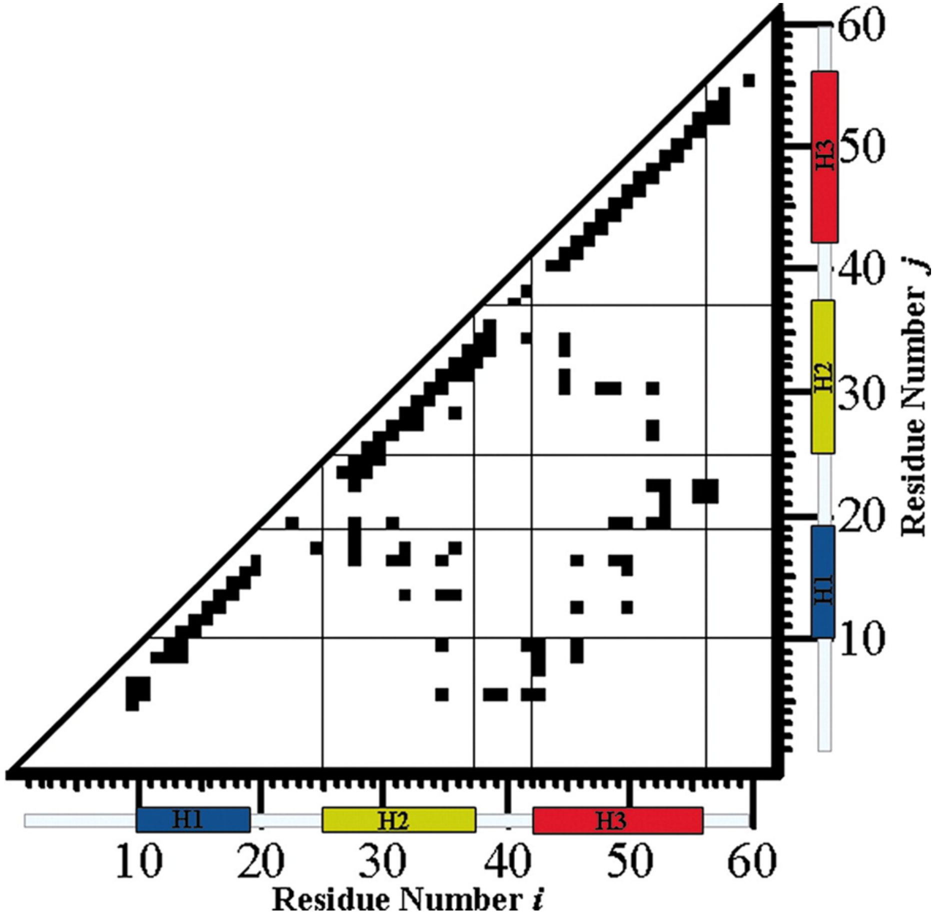
- columns do not mutate independently
- mutation in two columns are correlated, sites are near each other in space
    - source of structural information



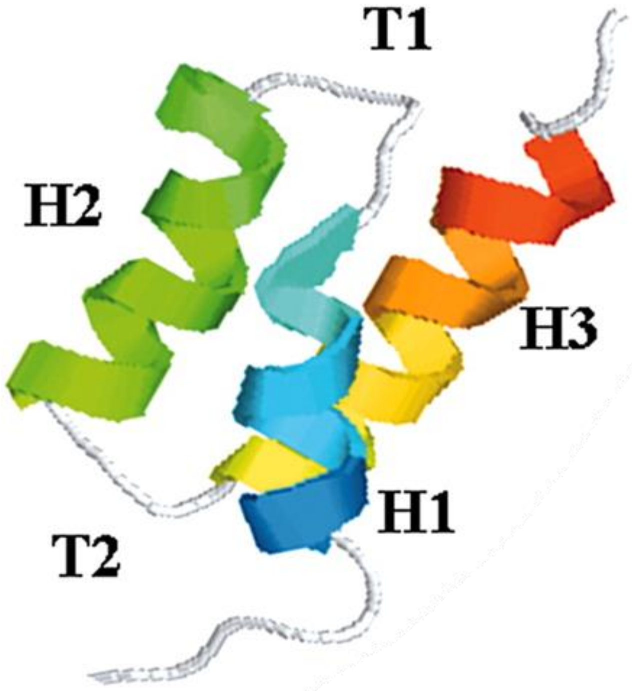


```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALEKMFLSFPTTKTYFPHFDLSHGSAQVKGHG
 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTHVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAYWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAHWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSAADKTNVKAGWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSAADKTNVKAFWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSADDKANIKAEWGKIGGHGAEYGAEALERMFCSFPTTKTYFPHFDVSHGSAQVKGHG
MLSPADKTNVKADWGKVGAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQG
VLSPADKTNVKACWGKVGAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQA
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKSNVKATWDKIGSHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHG
VLSPADKSNVKAWWGKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTGTYFPHFDLSHGSAQVKGHG
VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAHG
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAQWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSANDKSNVKAAWGKVGNHAPEYGAEALERMFLSFPTTKTYFPHFDLSHGSSQVKAHG
VLSPADKSNVKAAWGKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
    …       …        … …
```

from distances to contacts



distance geometry,
distance restraints,
model selection…



Itoh K , and Sasai M PNAS 2006;103:7298-7303

History

Idea from 80's or earlier*

- regular literature in 90's, 2000's
- little real success

Around 2010/2011 new ideas – methods

- Changes – will come to later

*Altschuh D, Vernet T, Berti P, Moras D, Nagai K (1988) Coordinated amino acid changes in homologous protein families. Protein Eng 2: 193–199.

How important is it ?

"epistasis, that is, instances when substitutions that are accepted in one genotype are deleterious in another" *

"we show that the observed dN/dS values and the observed patterns of amino-acid diversity at each site are jointly consistent with a non-epistatic model of protein evolution" **

*Breen, MS, Kemena, C, Vlasov, PK, Notredame, C., Kondrashov, FA, Nature, 490, 535-538 "Epistasis as the primary factor in molecular evolution"

** McCandlish, DM, Rajon, E., Shah, P., Ding, Y, Plotkin, JB, Nature, 497, E1, "The role of epistasis in evolution"

Alignments and noise

What is noise ?

- do all bad mutations disappear ?
  - what if there is $\frac{1}{100}$ chance of mutation being fixed ?
- biological weirdness / unusual environment
- sequencing errors

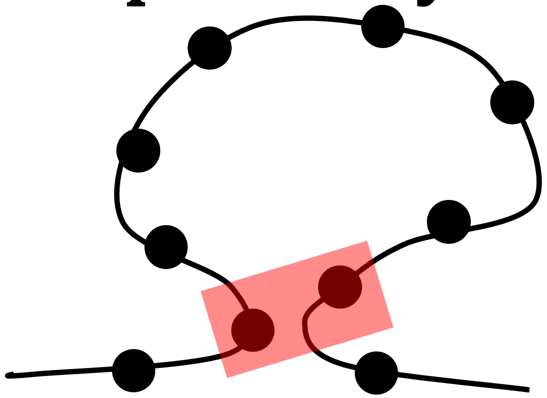
```
VLSPADKTNV
VLSPADKTNV
MLSPADKTNV
VLSPASKTNV
LVSPADKTNV
VLSPDDKTNV
...
```

Imagine we work with 500 – 10^3 sequences

Is it helpful ?

- bad news

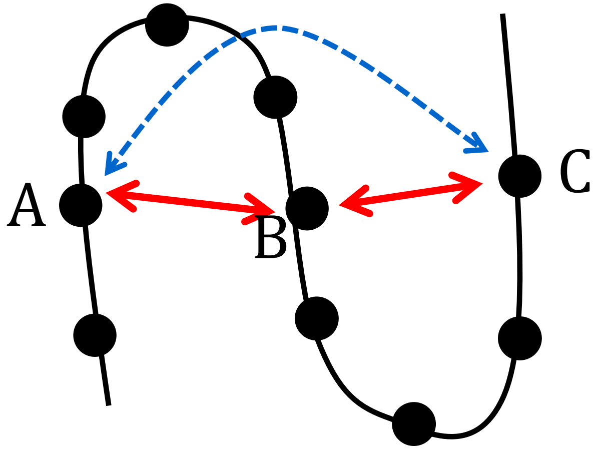
Does correlation mean proximity ?



```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALEKMFLSFPTTKTYFPHFDLSHGSAQVKGHG
 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTHVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAYWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAHWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSAADKTNVKAGWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSAADKTNVKAFWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSADDKANIKAEWGKIGGHGAEYGAEALERMFCSFPTTKTYFPHFDVSHGSAQVKGHG
MLSPADKTNVKADWGKVGAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQG
VLSPADKTNVKACWGKVGAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQA
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKSNVKATWDKIGSHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHG
VLSPADKSNVKAWWGKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTGTYFPHFDLSHGSAQVKGHG
VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAHG
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAQWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSANDKSNVKAAWGKVGNHAPEYGAEALERMFLSFPTTKTYFPHFDLSHGSSQVKAHG
VLSPADKSNVKAAWGKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
    …       …          …  …
```

Indirect effects

- A↔B ↔C
  A / C are correlated

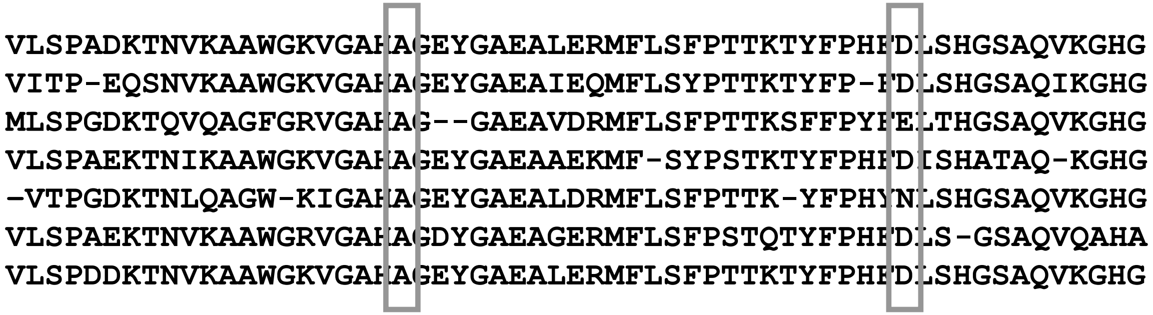


- Connected via substrate

Entropy / Information

normal entropy

$$S = -k \sum_{X}^{n_{states}} p_X \ln p_X$$



```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VITP-EQSNVKAAWGKVGAHAGEYGAEAIEQMFLSYPTTKTYFP-FDLSHGSAQIKGHG
MLSPGDKTQVQAGFGRVGAHAG--GAEAVDRMFLSFPTTKSFFPYFELTHGSAQVKGHG
VLSPAEKTNIKAAWGKVGAHAGEYGAEAAEKMF-SYPSTKTYFPHFDISHATAQ-KGHG
-VTPGDKTNLQAGW-KIGAHAGEYGAEALDRMFLSFPTTK-YFPHYNLSHGSAQVKGHG
VLSPAEKTNVKAAWGRVGAHAGDYGAEAGERMFLSFPSTQTYFPHFDLS-GSAQVQAHA
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

- forget k
- first column – no variation $S = 0$
- second .. $p_D = \frac{5}{7},\ p_E = \frac{1}{7},\ \ p_N = \frac{1}{7}$ so $S = -\left(\frac{5}{7}\ln\frac{5}{7} + \frac{1}{7}\ln\frac{1}{7} + \frac{1}{7}\ln\frac{1}{7}\right)$

Usual interpretation

- conservation

Other words

- how much information is present ?
- how good a predictor is this sequence for that sequence ?

I try to avoid using "information"
is it S, $-S$, $\log n - S$?

mutual information / entropy

- how much must certain pairs of amino acids be together ?
- amino acid types X and Y at sites i and j
- frequency (probability) of type X at site i is $p_{i,X}$
- frequency (probability) of pair XY at sites i and j is $p_{ij,XY}$
- mutual entropy (information)

$$I_{ij} = \sum_{X}^{n_{states}} \sum_{Y}^{n_{states}} p_{ij,XY} \ln \frac{p_{ij,XY}}{p_{i,X}\, p_{j,Y}}$$

n_{states} are the 20 amino acids

- why does it make sense ?

$$I_{ij} = \sum_{X}^{n_{states}} \sum_{Y}^{n_{states}} p_{ij,XY} \ln \frac{p_{ij,XY}}{p_{i,X}\, p_{j,Y}}$$

consider $\quad \ln \frac{p_{ij,XY}}{p_{i,X}\, p_{j,Y}}$

- how often would you expect to see X and Y together by chance ?
  - depends on the amount of X and Y

If there is no "mutual" information, $\frac{p_{ij,XY}}{p_{i,X}\, p_{j,Y}} = 1$ and $\ln 1 = 0$

- if they mutate independently, $I = 0$
- this measure says how much site i determines j (and vice versa)
- note summation over all XY pairs ..

Problems with mutual entropy

$$I_{ij} = \sum_X^{n_{states}} \sum_Y^{n_{states}} p_{ij,XY} \ln \frac{p_{ij,XY}}{p_{i,X}\, p_{j,Y}}$$

$\sum_X^n \sum_Y^n \ldots$

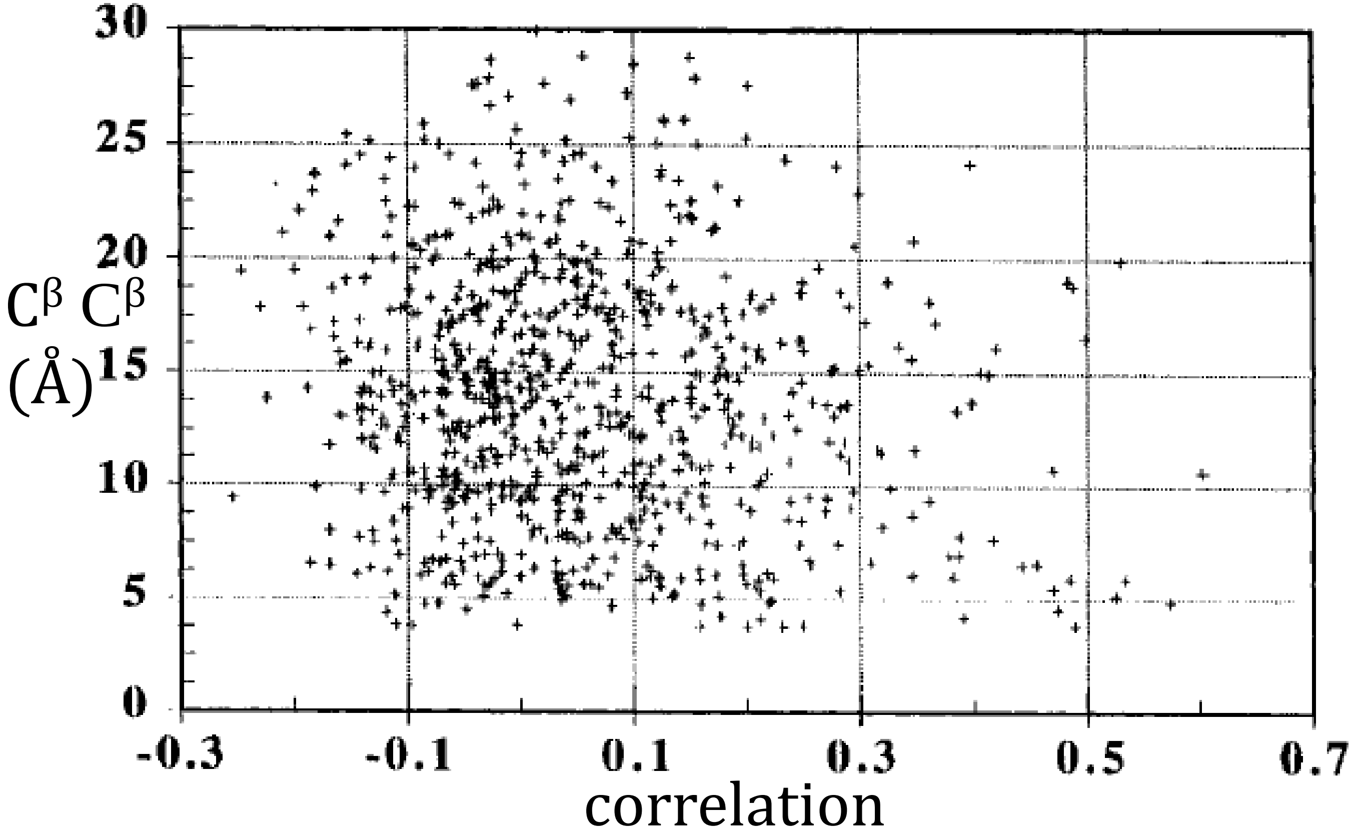
- $20 \times 20 = 400$ pairs
- need lots of data (1 000 sequences)
  - will encounter unusual sequences

Noise: What will most pairs be ?

- at most sites, many $p_X \approx 0$ (you do not find trp on surface or asp in middle)
  - if $p_X \approx p_Y \approx 0$ then $p_{i,X} p_{j,Y}$ very very small
  - the fraction $\ln \frac{p_{ij,XY}}{p_{i,X}\, p_{j,Y}}$ will be very sensitive to noise (unusual sequences)

Does it work ?

"predicted contacts in a small protein are fairly accurate"



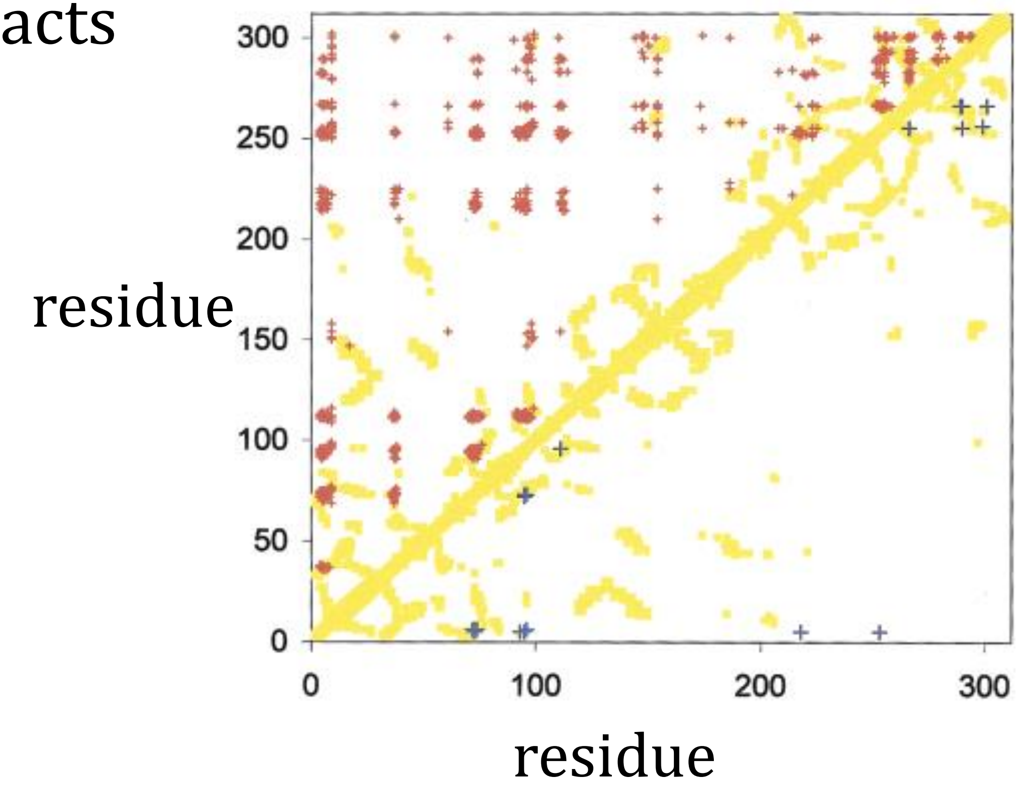
* Göbel, U, Sander, C, Schneider, R, Valencia, A, Proteins, 18, 309-317 (1994) Correlated mutations and residue contacts in proteins

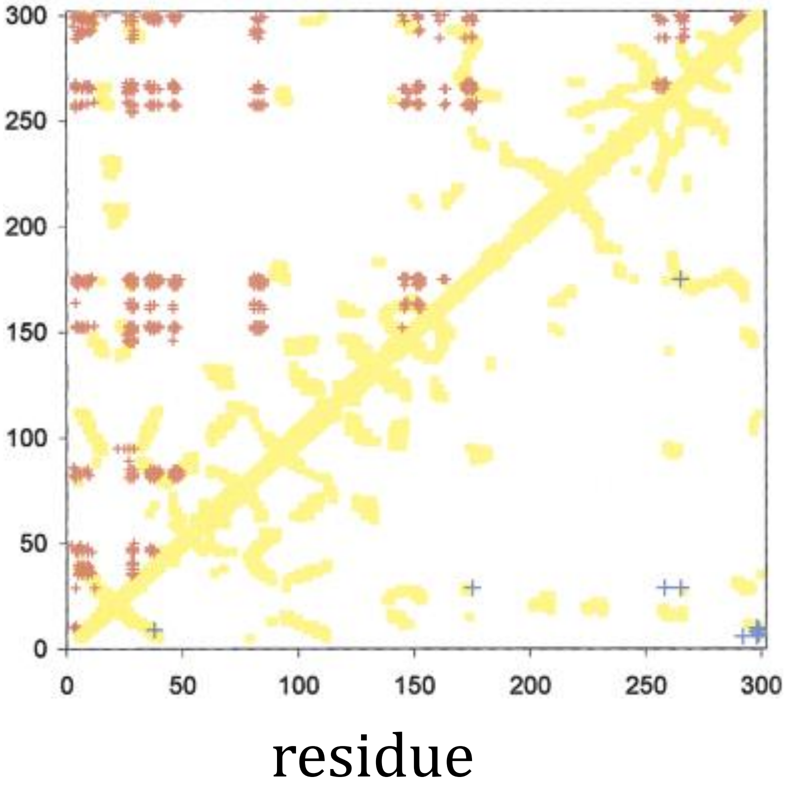
A few years later

Good show from two proteins

- red – predictions
- yellow – real contacts

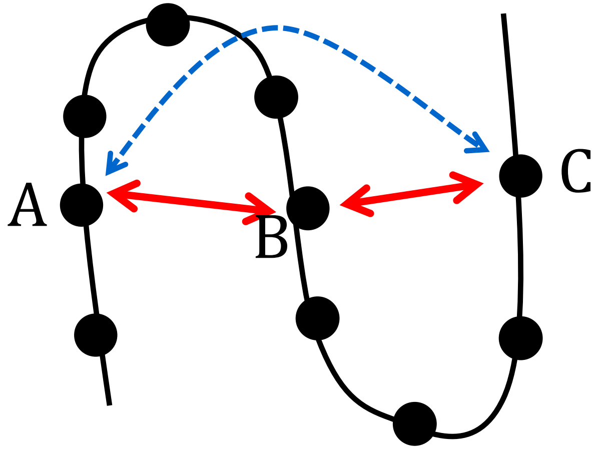
What has changed ?





Fariselli, P, Olmea, O, Valencia, A, Casadia, R, Proteins S5, 157-162, 2001, Progress .. inter-residue contacts of proteins..

transitive correlations



- transitive: A↔B↔C indirectly (transitively) A↔C

- Intuitive fix (will not work)
- visit all pairs of columns in alignment
- make list of correlated pairs

- sort list
- use n most correlated pairs

- why will it not work ?

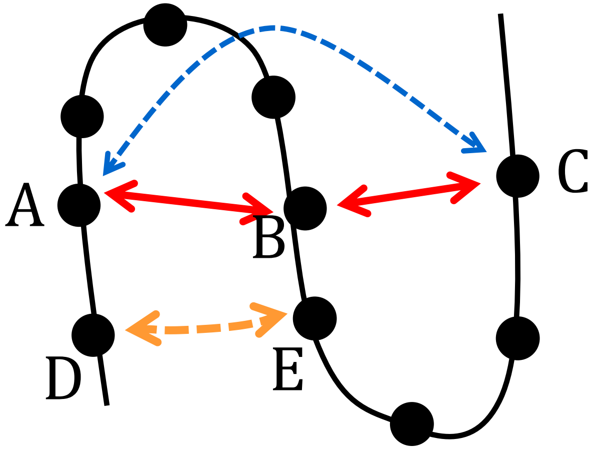
Simple fix does not work

imagine D is on surface

  - varies a lot
  - swaps asp↔glu or ser↔thr

- cross correlation DE is weaker than AC
- DE will be removed before the transitive relation (AC)

AB
BC
AC
DE



Residue similarities

- asp/glu, asn/gln, ser/thr, ile/leu, …

- The sorted list will only be a weak indicator of how direct relations are

The statistical problem

Earlier

$$I_{ij} = \sum_{X}^{n_{states}} \sum_{Y}^{n_{states}} p_{ij,XY} \ln \frac{p_{ij,XY}}{p_{i,X}\, p_{j,Y}}$$

- assumes that residues and pairs are independent of the sequence they are in…

  `ABCIEFGIJKLM`

- but **`I`** depends on **`ABC-EFG...`** and **`I..J`** on **`ABC-EFG..`**
- this effect is not small
- can one account for background distributions ?
  - properly ?
    - too expensive
  - approximations..

covariance

Principle problem .. our $p_{X,i}$ and $p_{XY,ij}$ do not account for background (rest of sequence)

- treat in an average manner

What would you expect if everything was independently distributed ?

$p_{XY,ij} = p_{X,i}\, p_{Y,j}$ or $p_{XY,ij} - p_{X,i}\, p_{Y,j} = 0$

- difference from what you expect is the key.. define a covariance matrix

$C_{ij} = p_{XY,ij} - p_{X,i}\, p_{Y,j}$

Huge difference to earlier version

- before $I = \sum_X^{n_{states}} \sum_Y^{n_{states}} p_{ij,XY} \ln \frac{p_{ij,XY}}{p_{i,X}\, p_{j,Y}}$ one number for pair of columns i, j
- now matrix C_{ij} … more informative, but not so practical

from matrix to single number – example philosophy

several approaches (details not for exam)

if C tells me how objects move together

C^{-1} tells me about the couplings

Here

- C_{ij} tells me how amino acid types in columns i, j move together (from expected values)
- C_{ij}^{-1} tells me how they are coupled (elements tell me about specific amino acids)
  - if columns move independently C_{ij} will not have off-diagonal elements
- if C_{ij}^{-1} has lots of non-zero elements, there are lots of couplings

- Primitive – sum up the elements of C_{ij}^{-1}
- sounds better: use ℓ_1 norm coupling/contact $= \sum_X^{20} \sum_Y^{20} \left|\Theta_{ij}^{XY}\right|$ where Θ comes from C_{ij}^{-1}

summarise the steps and ideas

- mutual entropy sounds good, does not account for dependencies on whole sequence
- covariance matrix approach much much better
  - remember idea of $p_{XY,ij} - p_{X,i}\, p_{Y,j}$
- need some way to go from covariance matrices to estimates of connections between columns in multiple alignment

- does it all work ?

from contacts to structure

Most obvious route

- extract contact predictions

Then

- use as C$^\beta$ C$^\beta$ restraints – distance less than 8 Å

maybe

- use as restraints in an MD simulation

or

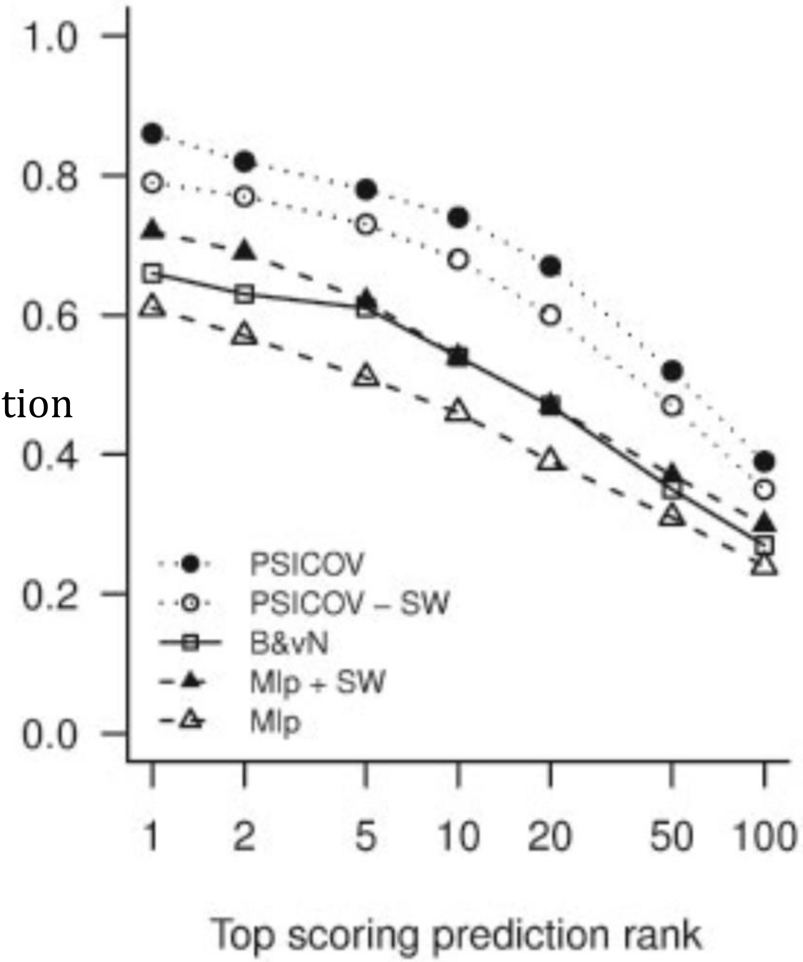
- use speculative fold recognition method and see which answers are plausible

Consider how many predicted contacts seem to be correct

150 proteins

- predicted contacts
- rank by confidence
- compare with known structures

- another group showing contacts on structure …



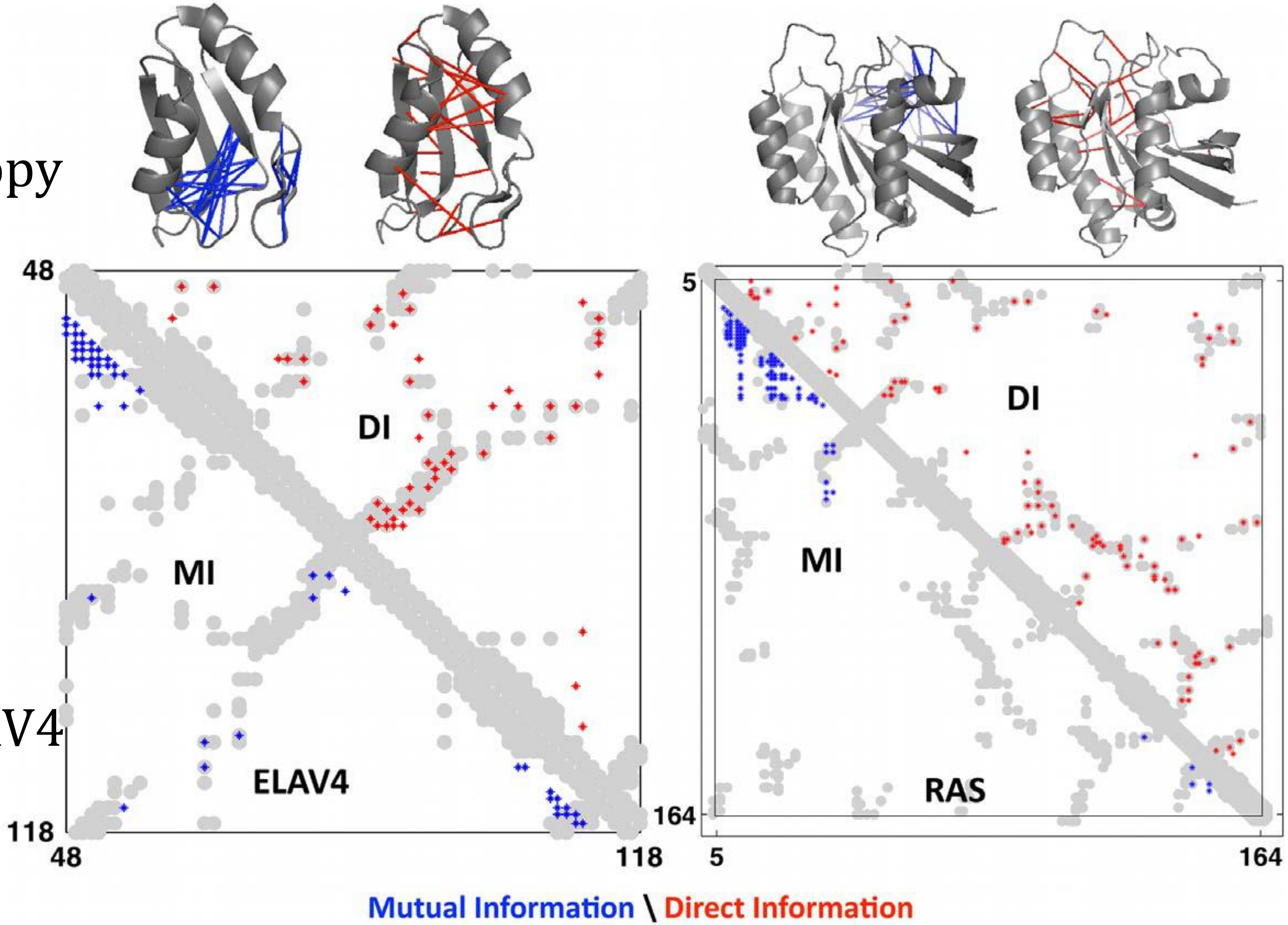
precision
(positive prediction
value)

Jones, DT, Buchan, DWA, Cozzetto, D, Ponti, M, Bioinformatics, 28, 184-190, … precise structural contact prediction..

- contacts from mutual entropy
  blue
- based on covariance
  red

- correct contacts
  grey

- mapped on to RAS and ELAV4



Marks, DS, Colwell, LJ, Sheridan, R, Hopf.. Sander, C. PLOS One, 6, e28766, 2011, Protein 3D … sequence variation

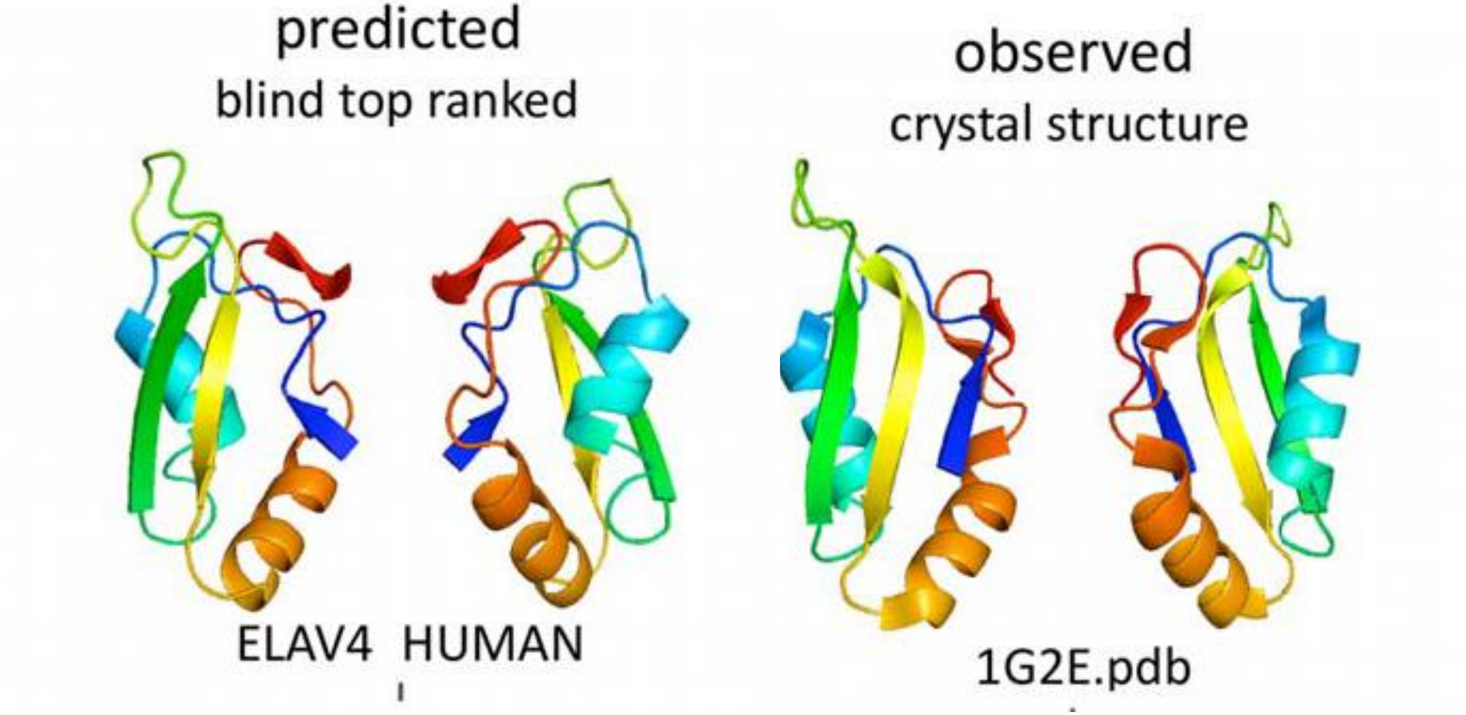
calculating structures

Method

- contacts from multiple alignment
- secondary structure prediction
- distance geometry + refinement

- + more examples

- looks too good



Marks, DS, Colwell, LJ, Sheridan, R, Hopf.. Sander, C. PLOS One, 6, e28766, 2011, Protein 3D … sequence variation

Is the problem solved ?

To come..

- how many sequences ?
- noise
- proteins to apply it to
- phylogenetic affects / sampling

How many sequences ?

Two examples

- 500 to 74×10^3 choose by some criterion of similarity
- 10^3 chosen arbitrarily

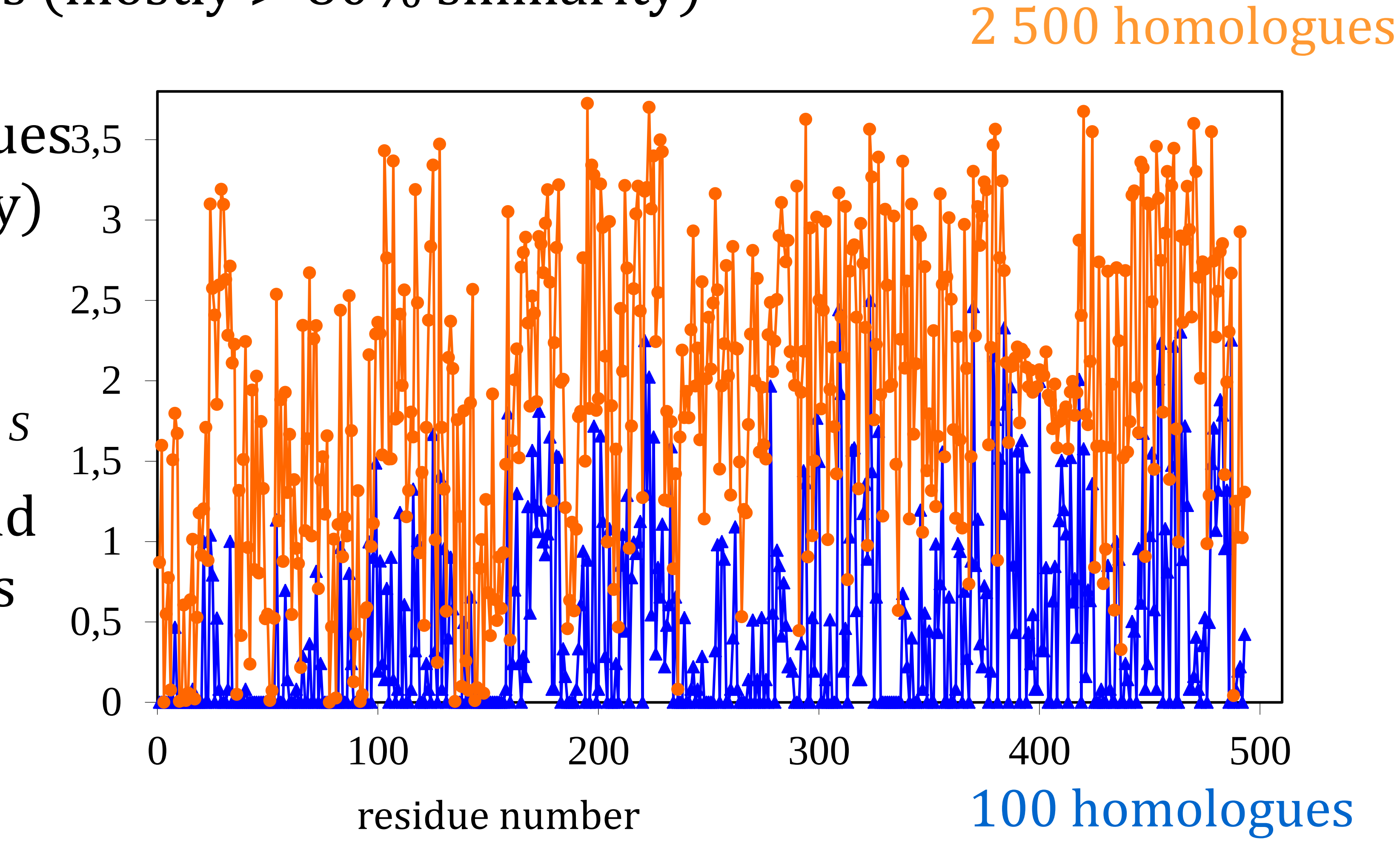
- see the importance by just looking at entropy

Entropy and number of homologues

Example sequence (1ab4, DNA gyrase)

- find 100 close homologues (mostly > 80% similarity) – calculate conservation
- find 2 500 close homologues (mostly > 50 % similarity)
- calculate conservation

- how many changes depend on how many homologues you have



Noise

- unusual sequences, errors, unusual environments

Evolution

- random events with some selection
- if I have many many random parameters some will always appear coupled

- I find a p-value of 10^{-3} must it be significant ?
- what if I look 10^{5} times ?

Applicability

Does the method really work ?

- nobody knows

Applications in literature

- 1000s of homologues
- usually a crystal structure was solved – use modelling

Phylogenetic and sampling effects

In an alignment column you see

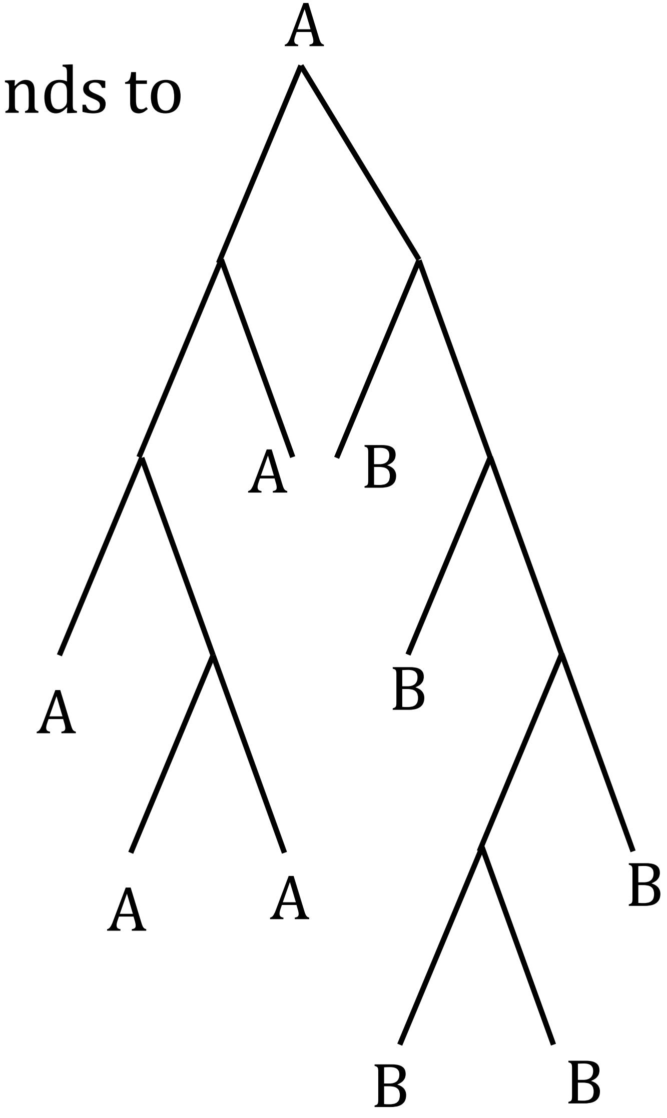
- appears to be random A/B

```
A
B
A
A
B
A
A
B
B
```

In tree

- one mutation only
- looks like a high information site

corresponds to

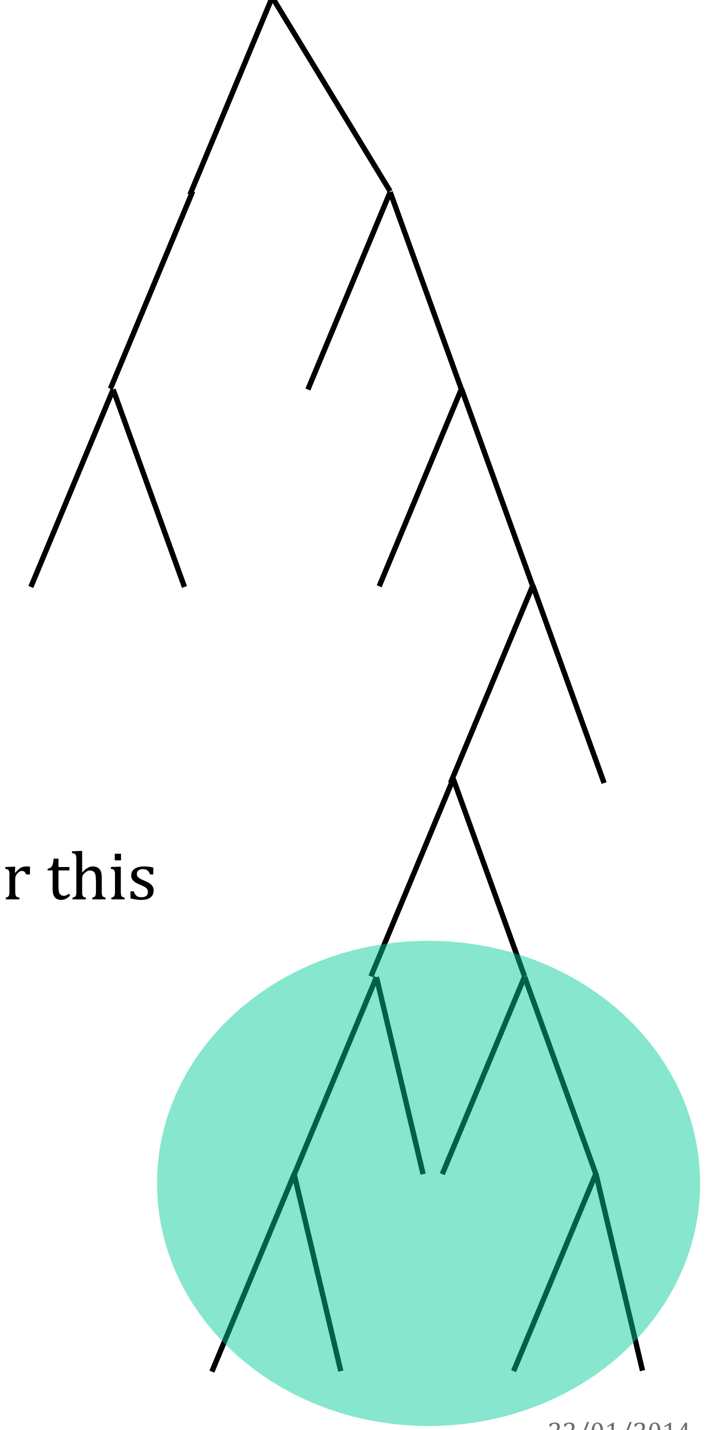


Sampling

Not even across nature

- green area
  - "late radiation" ? (evolution)
  - some clinical bacterium
    - important
    - cheap to sequence

- the practical schemes use *ad hoc* methods to account for this



summary

Correlated mutations – long history

- much promise in last 3 – 5 years

Mutual information/entropy methods vs covariance

- transitive versus direct relationships

Problems

- how many homologues
- noise
- phylogenetics / sampling
- need lots of data

- not proven on unknown cases