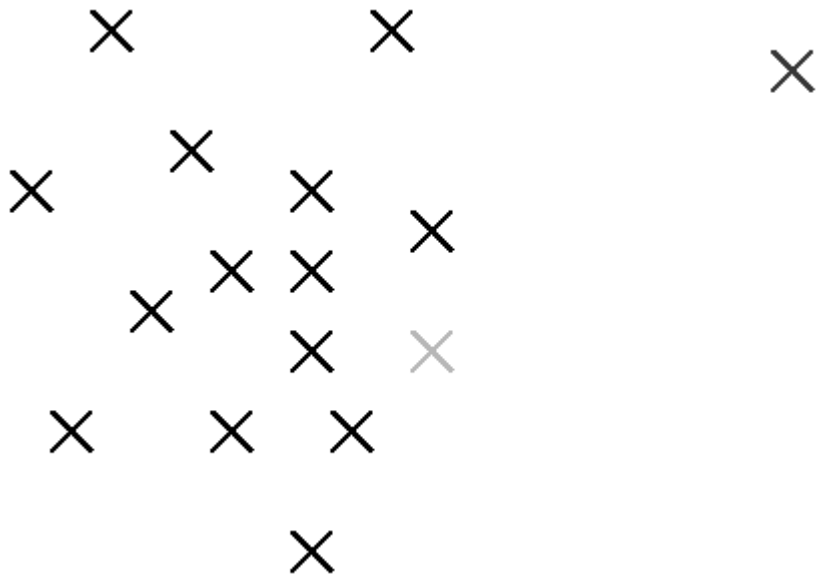
Bioinformatics
(3 lectures)

Thomas Huber
Supercomputer Facility
Australian National University
Thomas.Huber@anu.edu.au

- Why bother about proteins/prediction
- What is bioinformatics
- Protein databases
- Making use of database information
  - Predictions
- Protein Design

What is Bioinformatics?

- Handling lots of information
  - Concentrate knowledge
    - public databases
  - Summarise knowledge in principles
    - knowledge acquisition (data mining)
  - Apply principles
    - predictions



Why do we care about Protein Structures/ Prediction?

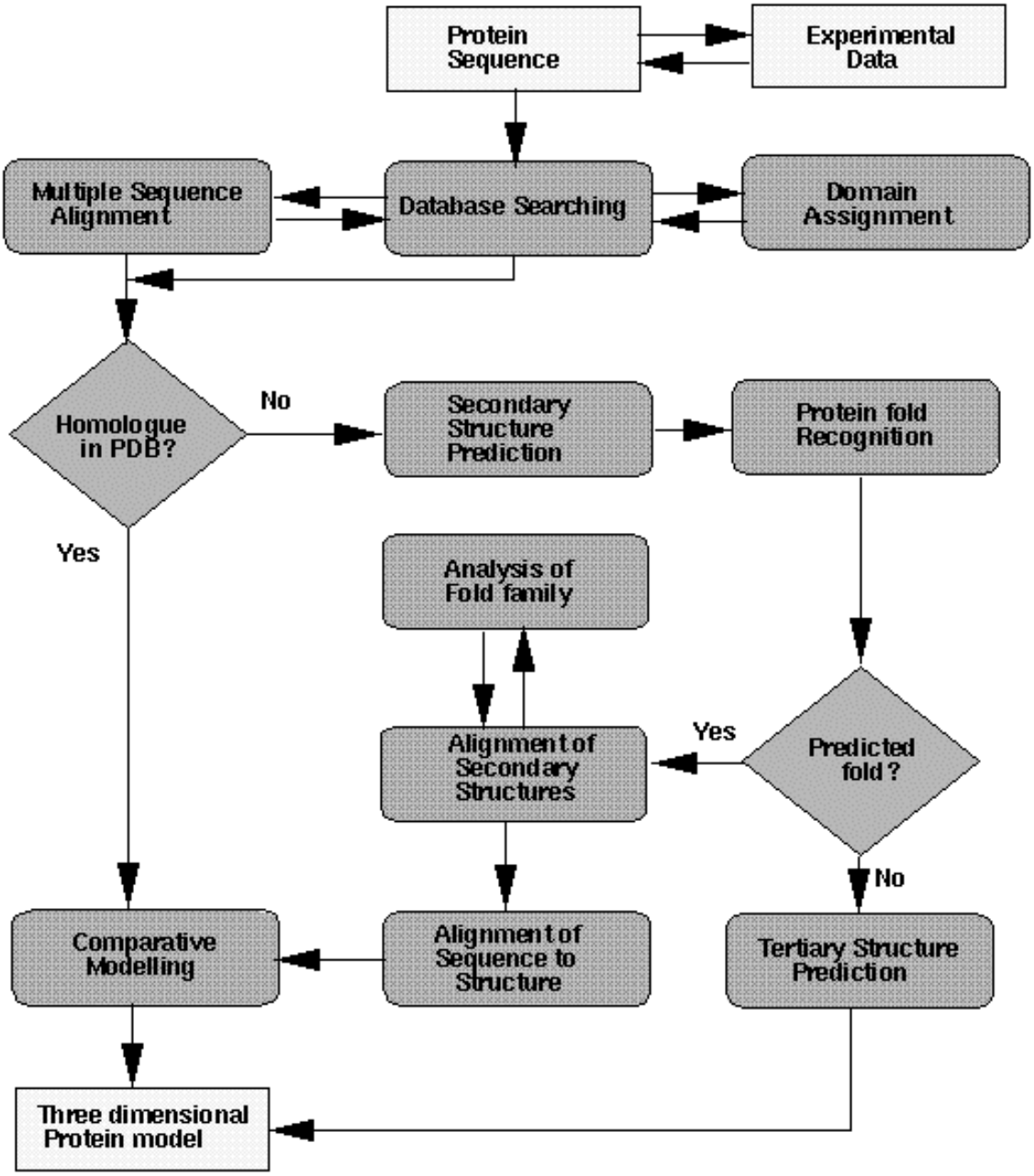
- Academic curiosity?
  - Understanding how nature works
- Drug & Ligand design
  - Need protein structure to design molecules which inhibit/excite
    - cure all sorts of diseases
- Protein design
  - making better proteins
    - sensor proteins
    - industrial catalysts (washing powder, synthetic reactions, …)
- Urgency of prediction
  - ≈10000 structures are determined
    - insignificant compared to all proteins
  - sequencing = fast & cheap
  - structure determination = hard & expensive

Protein Databases

- Collection of protein information
  - cunningly organised
    - cross references
    - easily accessible

- Different information = different databases
  - Literature databases (Medline)
  - Sequence databases (Swissprot)
  - Pattern (finger print) databases (Prints)
  - Structure databases (PDB)
  - Function databases (PFMP)

Prediction of Protein Structure



Sequence Search

- Sequences are major source of biology
  - access to ≈85000 annotated sequences
  - much more to come from DNA sequencing

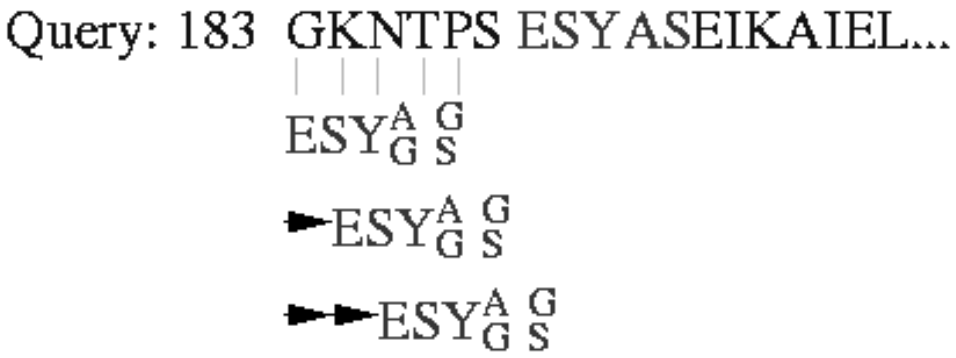
- What information to look for?
  - Sequence pattern
    - many protein families have sequence "finger prints"
  - Similar sequences:
    - Observation: Two proteins with sequence identity >35% adopt same structure
    - Family of sequences ⇒ useful for structure prediction

Searching Sequence "Finger Prints"

- What are protein "finger prints"?
  - a pattern of conserved residues (often with functional importance)
  - unique (or highly specific) for a protein family
  - e.g. Carboxypeptidases finger print

    [LIVM]-x-[GTA]-E-S-Y-[AG]-[GS]

- Searching for finger prints



Sequence Alignment

- What is a similar sequence?

```
Query: 183 GKNTPSMCLEAEIKAIELAVKVSLDYL 209

Sbjct: 181 GQVPPSMCYEMELEAVKVAIEVALEEL 207

           score = ?
```

  - With finger prints: Yes/No
  - Sequence similarity (≈1gozillion measures)
    - identity: score 1 if residues are the same
      score 0 if residues are different

```
Query: 183 GKNTPSMCLEAEIKAIELAVKVSLDYL 209
           G   PSMC E E  A   A  V L  L
Sbjct: 181 GQVPPSMCYEMELEAVKVAIEVALEEL 207

           score = 12
```

    - physico-chemical (e.g. positives, hydrophobicity):

```
Query: 183 GKNTPSMCLEAEIKAIELAVKVSLDYL 209
           G+  PSMC E E++A+++A++V+L+ L
Sbjct: 181 GQVPPSMCYEMELEAVKVAIEVALEEL 207

           score = 22
```

Evolutionary Similarity

- PAM (Probability of Accepted Mutation)
  - Align sequences with >85% identity
  - Reconstruct phylogenetic tree
  - Compute mutation probabilities for 1 PAM of evolutionary distance
  - Calculate log odds

$$S_{i \leftarrow j} = \log\left(\frac{p_{i \leftarrow j}}{p_i}\right)$$

$p_{i \leftarrow j}$ = probability amino acid j was replaced by i

p_i = probability of occurence of amino acid i

  - extrapolate matrices to desired evolutionary distance
    - e.g. PAM250 for evolutionary distant sequence

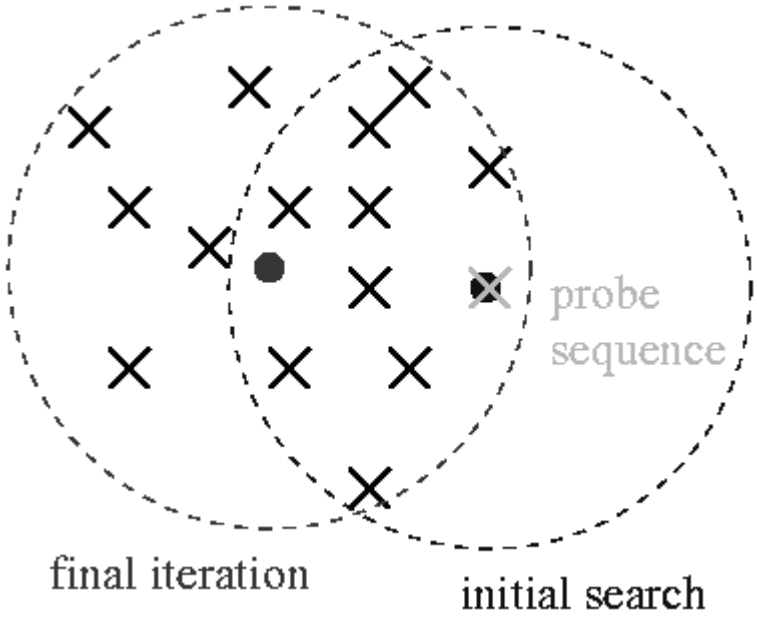
Searching for Similar Sequences

- What is the difference to searching for finger prints?
  - Gaps and insertions: nasty complication

```
Query: 6    GNVNR---IDTTGASCKTAKPEGLSYCG
            GN+ +   + +T +S  T    G SY G
Sbjct: 1393 GNITQYGIVTSTTSSGGTPSSTGGSYSG
```

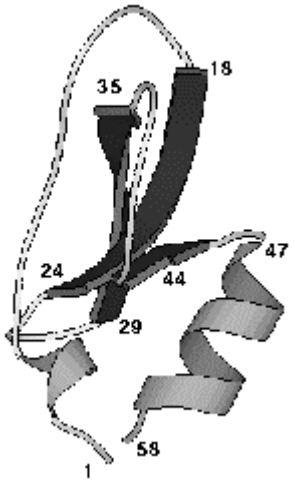
Finding Distant Homologues

- Iterative sequence alignment (✈-Blast)



Predicting Secondary Structure

- Secondary structure (a reminder)
  - simple (but not sufficient) description of structure




```
...,....1...,....2...,....3...,...
RPDFCLEPPYTGPCKARIIRYFYNAKAGLAQTFVYGGC
  HHHH           EEEEEEE    EEEEEEE
```

- Prediction of secondary structure
  - relation of protein sequence to structure
  - statistically based prediction
  - pattern based prediction

Statistical Based Prediction

- Amino acids have preferences for secondary structure

Odds preferences of amino acids from a set of 600 non-redundant proteins (87000 aa)

Amino Acid	✂	〰	other
ALA	1.472	0.780	0.784
GLU	1.385	0.745	0.862
LEU	1.352	1.123	0.696
GLN	1.332	0.789	0.877
MET	1.290	0.978	0.811
ARG	1.245	0.892	0.885
LYS	1.161	0.828	0.975
VAL	0.894	1.806	0.672
ILE	1.020	1.712	0.632
TYR	0.974	1.466	0.786
PHE	0.962	1.417	0.819
TRP	0.989	1.271	0.873
THR	0.759	1.245	1.044
CYS	0.748	1.209	1.070
PRO	0.409	0.455	1.678
GLY	0.444	0.644	1.560
ASP	0.862	0/547	1.320
ASN	0.799	0.671	1.302
SER	0.771	0.866	1.225
HIS	0.922	1.035	1.037

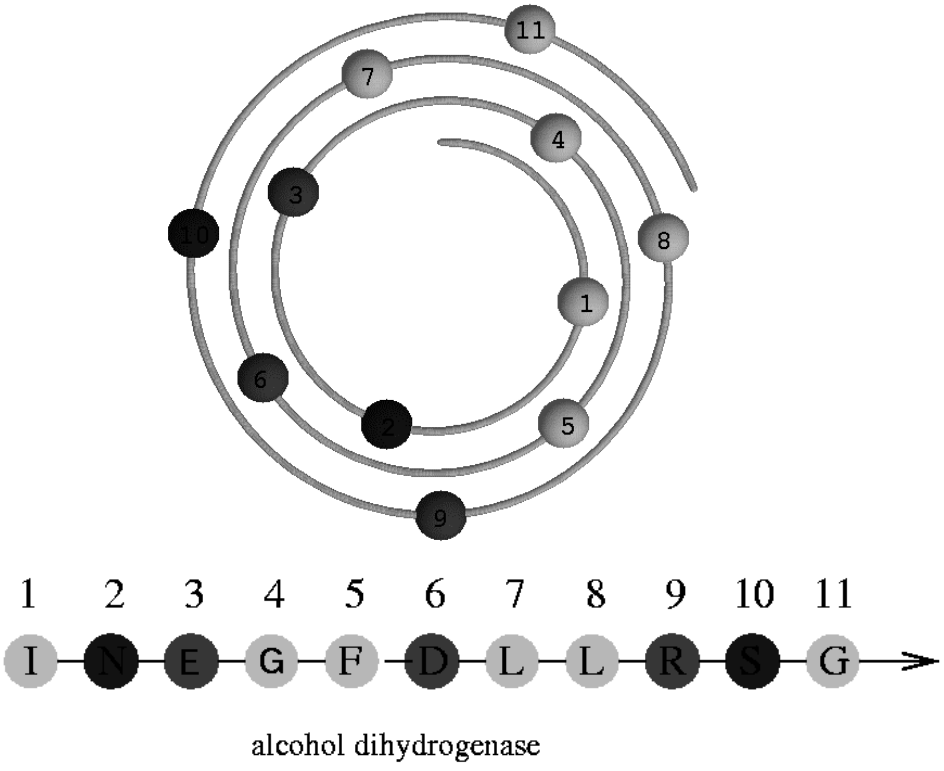
- What are the odds?

$$p_\alpha^i = n_\alpha^i / n_\alpha^i \text{(by chance)}$$

$$n_\alpha^i \text{(by chance)} = n_\alpha \left(n^i / n^{\text{tot}} \right)$$

Pattern Based Prediction

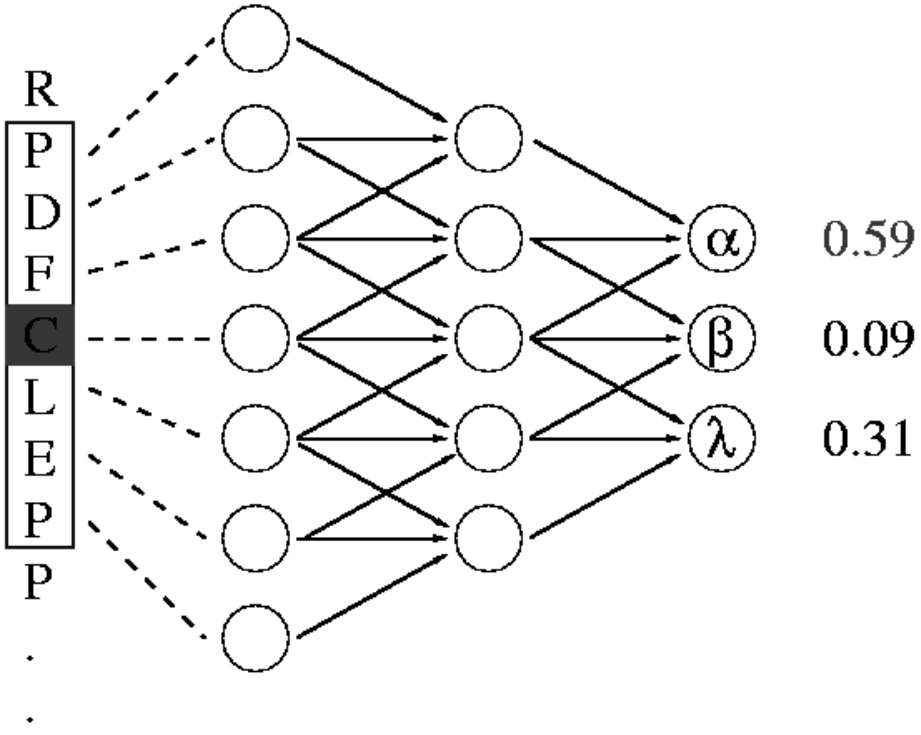
- Do amino acid pattern exist?
  - Yes but the code is not always obeyed
    - Same sequence of 5 residues is sometimes in ✂-helix and at other times in 🖋-strand
- BUT pattern have high preferences
- A good predictor: The helical wheel
  - Helices are likely on outside of proteins
  - I, I+3 and I+4 hydrophobic interface



alcohol dihydrogenase

Prediction with Neural Networks

- Not enough statistic for all pattern
  - for 5 residues 20^5 ($3.2*10^6$) pattern
- How to reduce the number of parameters?
  - Train a neural network to “learn” to predict secondary structure



How Accurate are the Predictions?

- Secondary structure prediction is not accurate
  - random prediction

    ≈33% correct
  - simple preference based predictors:

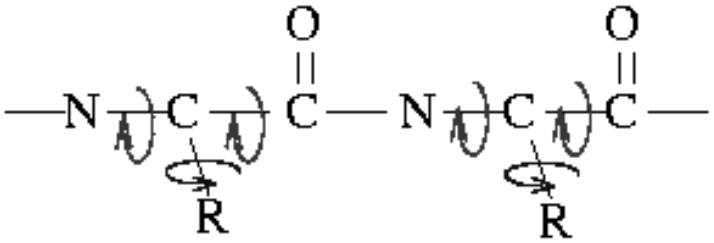
    ≈55% correct
  - pattern based predictors:

    up to ≈65% correct
  - best neural network based predictors using families of homologous sequences:

    ≈70-73% correct

Prediction of 3D Structure

- ab initio prediction
  - much too hard
    - number of possible conformations = astronomical



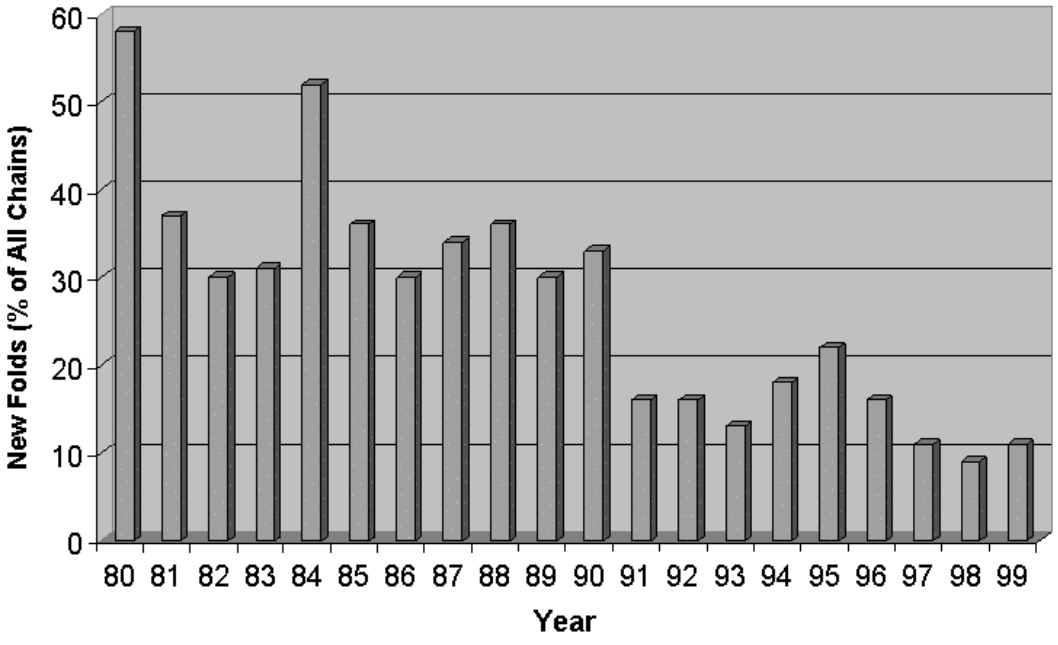
- 3 possible rotamers per dihedral angle
- 2 dihedral angles per amino acid

$\Rightarrow$ for protein with 100 residues

3^{100} possibilities

Fold recognition

- More moderate goal:
  - recognise if sequence matches a protein structure
- Is this useful?
  - $\approx 10^4$ protein structures determined
  - $< 10^3$ protein folds

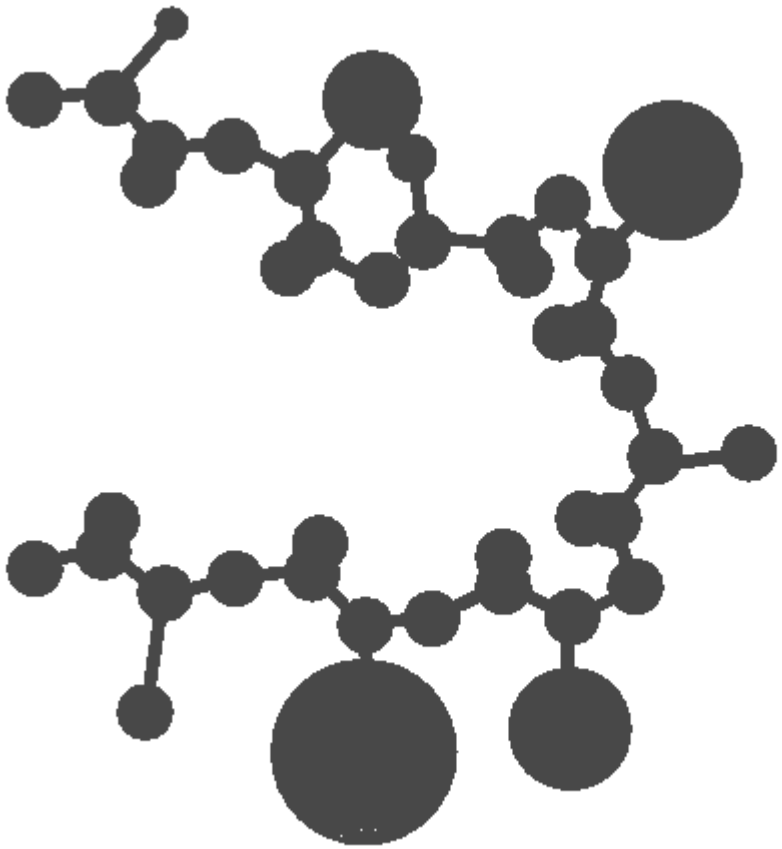


How Fold Recognition Works

- Finding a match in a structure disco

What is a match?

- Calcululate happiness of pair
  - similar to energy in molecular modeling
    - interactions between all pairs of residues
  - captures amino acid preferences
    - BUT not necessarily physics



backbone +
1 sidechain

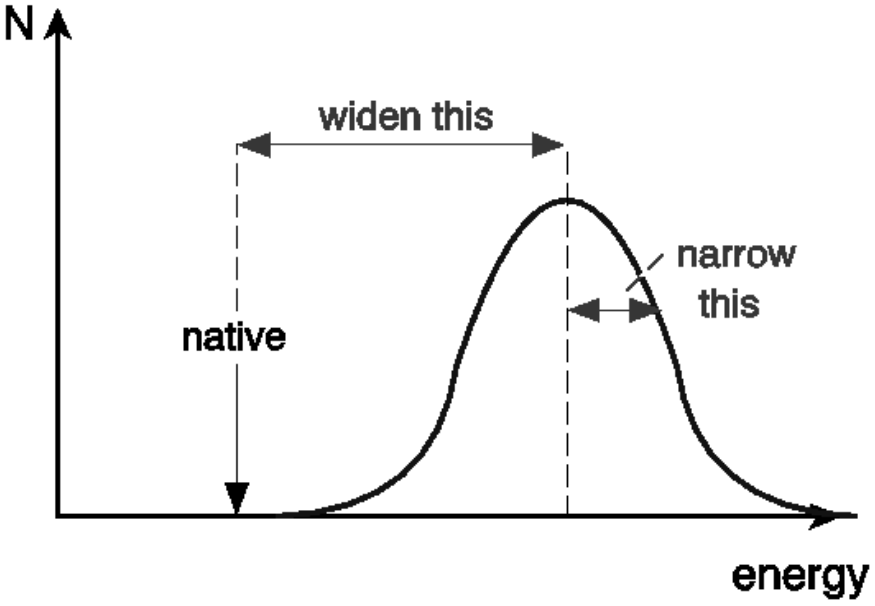
Scoring Schemes

- Plentiful like sequence similarity matrices
  - log odds (Boltzman based force fields)

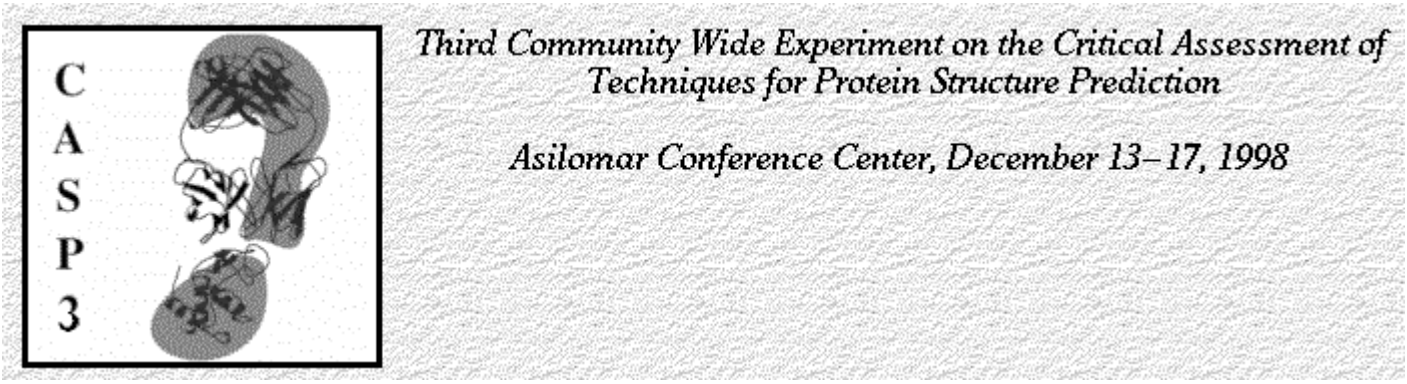
$$\text{score} = \sum_{j>i}^{N} \log(p(s_i, s_j, d_{ij}))$$

$$p(s_i, s_j, d_{ij}) = n(s_i, s_j, d_{ij}) / n(s_i, s_j, d_{ij}) \text{(by chance)}$$

    - c.f. Boltzman's law $\quad p = \exp\left(-\frac{E}{k_B T}\right)$
  - optimised for discrimination



How Successful?



- Blind test of methods (and people)
  - methods always work better when one knows answer
- ≈30 proteins to predict
- ≈90 groups
- Best groups: ≈25% (partly) correct

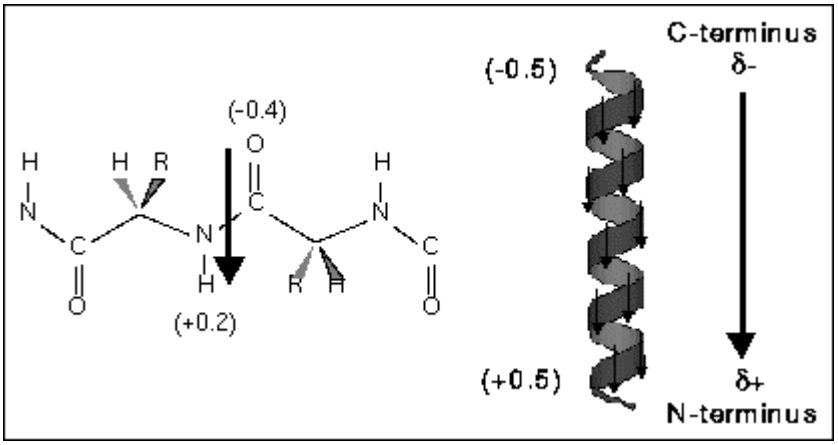
  BUT
  - accuracy (probably) not good enough to be useful for X-ray structure determination

Protein Design

- The Inverse Problem
  - Is there a better sequence match for a structure?
- What is “better”?
  - More stable
  - Better function
- Why important?
  - Many industrial applications
    - E.g. enzymes in washing powder
      - should be stable at high temperatures
      - work faster at low temperature
      - …

Rational Approaches For More Stable Proteins

- Rules of thumb (work nearly always)
  - Restriction of conformational space
    - Covalent bonds between close residues
      - e.g. disulfide bonds
    - Rigid residues
      - e.g. proline instead of glycin
  - Introducing favourable interactions
    - salt bridges
    - compensating for helix dipol



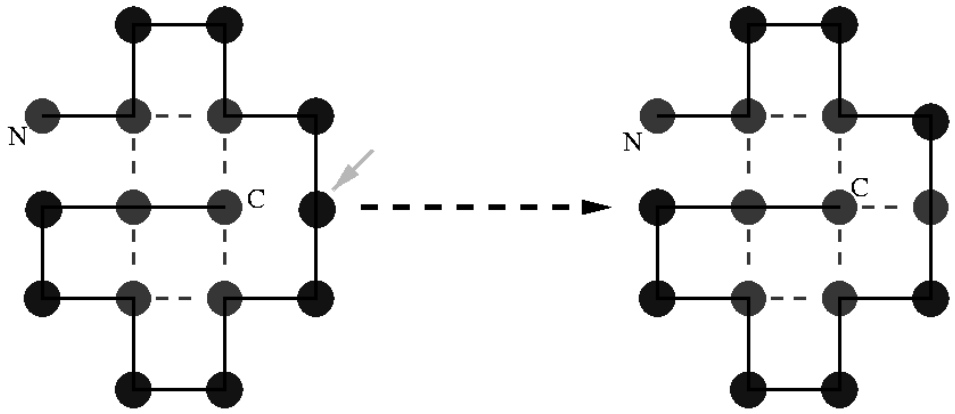
Naïve Approach

- Use happiness score
  - e.g. score from fold recognition
- Change sequence to increase happiness

Why Naïve?

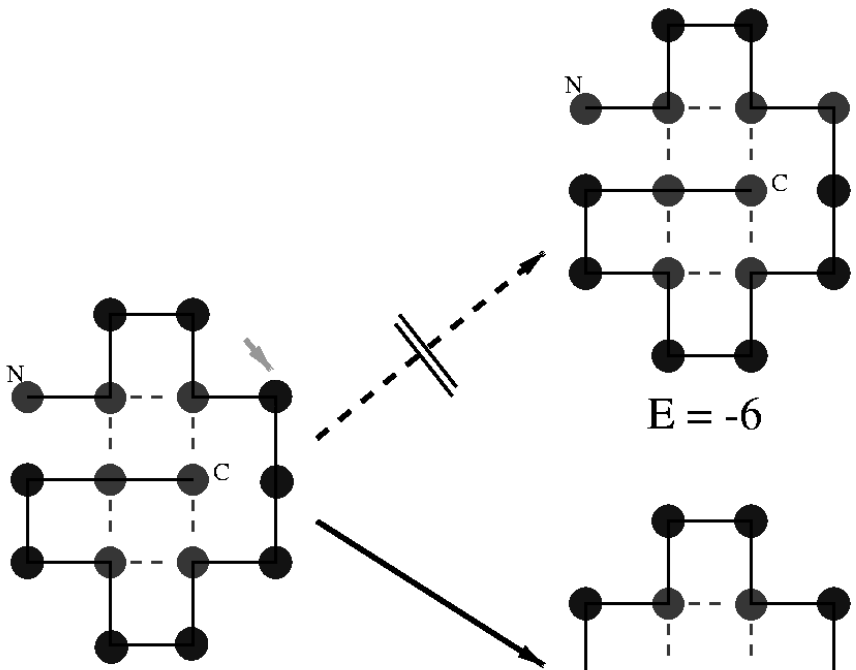
- Stability = difference between folded and unfolded state
- Aim:
  - Increase gap of happiness
  - NOT absolute happiness

Pitfalls



E = -6
F = -0.65981

E = -7
F = 0.00005



E = -6

E = -6
F = -0.65981

E = -7
F = -0.68215

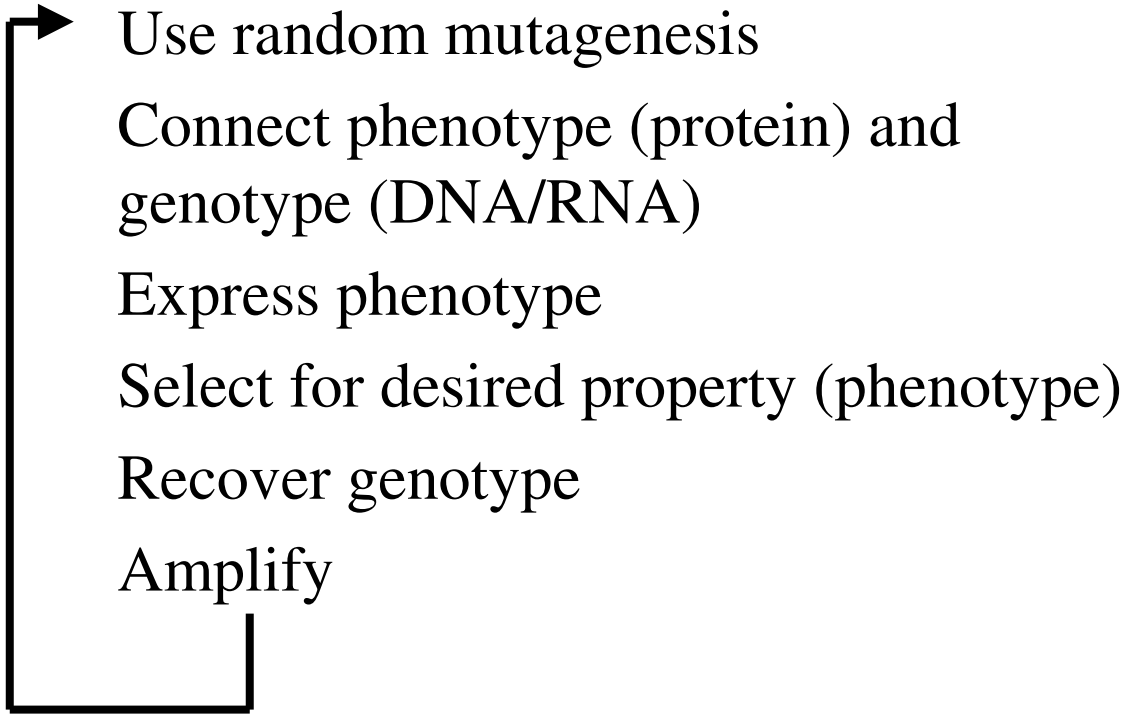
Combinatorial Design (Experimental)

- Basic Idea
  - Generate large number of sequence variations
  - Select pool for desired property

- Peptide libraries
  - systematic synthesis
    - (e.g. all tri-peptides)
      - expensive
  - mix & code

Directed Evolution Techniques

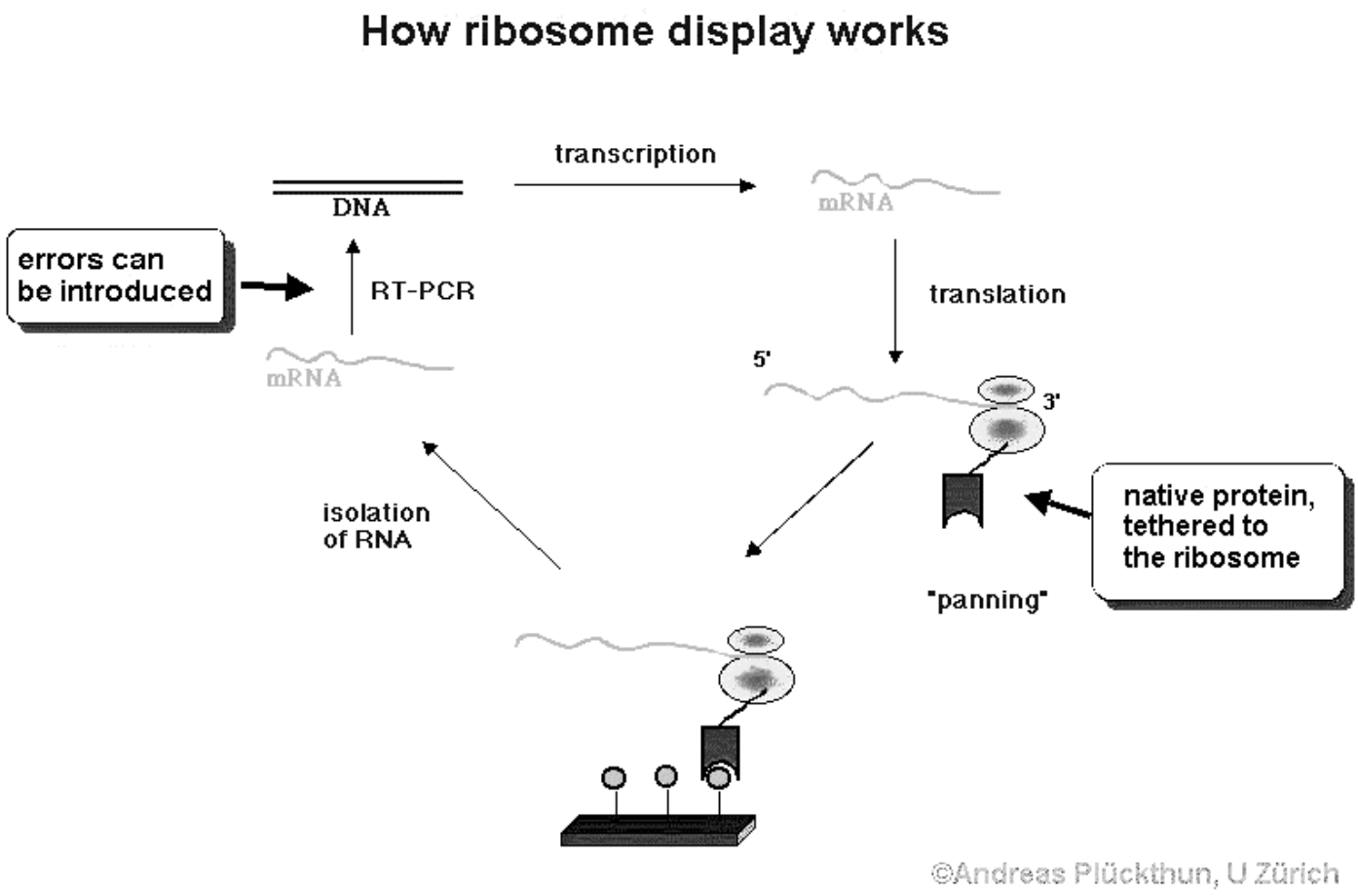
- Idea



- Where is genotype and phenotype connected?
  - In Viruses (coat protein/virus DNA)
  - At Ribosome

Phage Display

Ribosomal Display



- Advantage:
  - much bigger library (10^{12}-10^{13} copies)
- Problems:
  - How connect RNA with Ribosome?
  - How connect Protein to Ribosome?

Summary

- Protein databases = huge collection of knowledge
- Bioinformatics = making use of this knowledge
- Simplest way to extract knowledge = statistical based
  - log odds
- Structure prediction = interpolation of rules (extrapolation is dangerous)
- Protein design industrially important
  - rational design not yet come to age
  - combinatorial design = very powerful
    - accelerated spiral of information (hopefully knowledge)