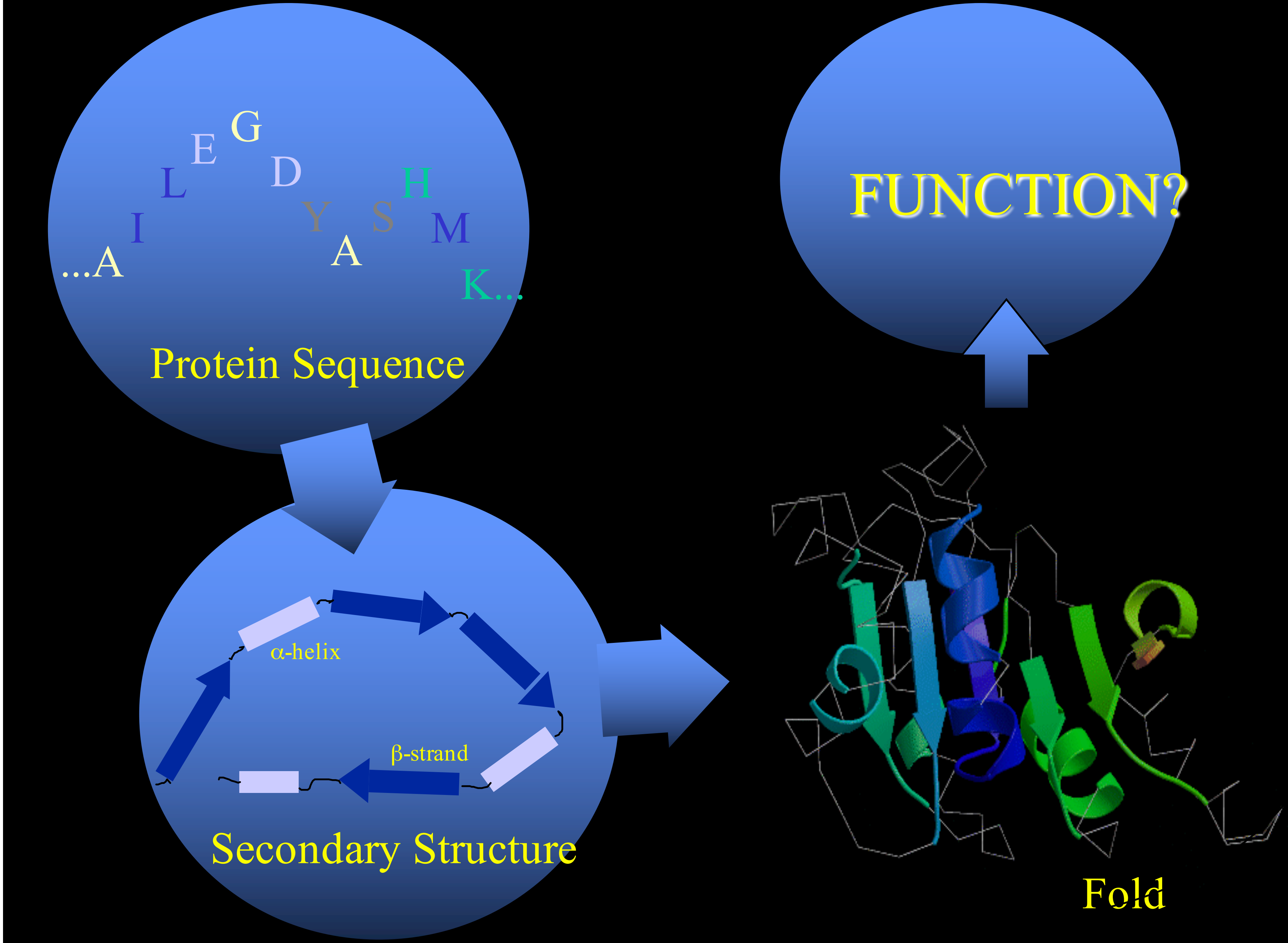
Protein Structure Prediction

Geoff Barton

www.compbio.dundee.ac.uk



...A I L E G D Y A S H M K...

Protein Sequence

α-helix

β-strand

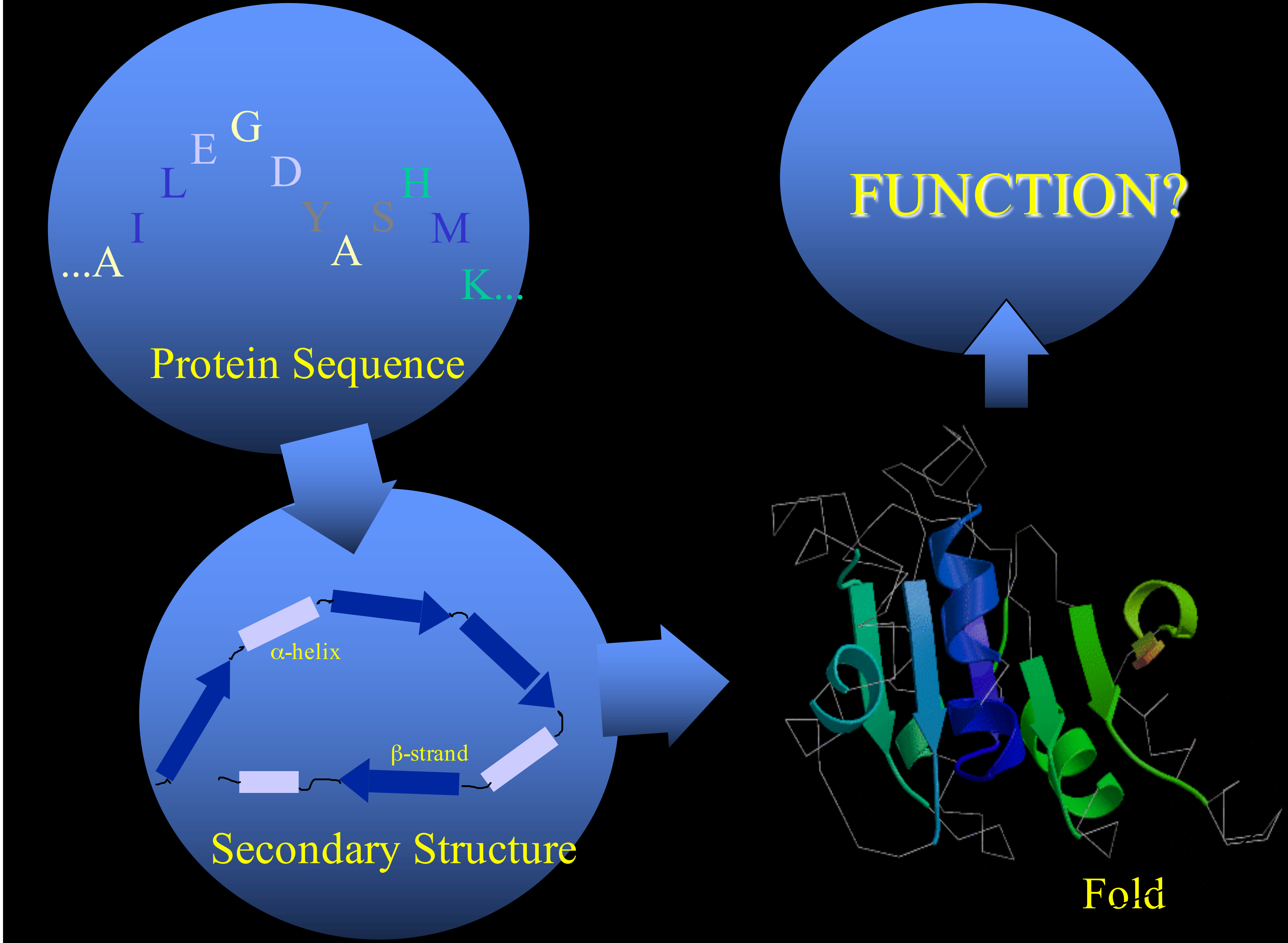
Secondary Structure

Fold

FUNCTION?

Protein Structure Prediction

- Since all the information about the protein three-dimensional structure is contained in the protein sequence, it should be possible to predict the protein three-dimensional structure given only the sequence…
- However, this turns out put be pretty hard to do…



...A I L E G D Y A S H M K...

Protein Sequence

α-helix

β-strand

Secondary Structure

Fold

FUNCTION?

Before talking about prediction…

Protein Structure Determination by Experiment

- From X-ray crystallography
- From NMR spectroscopy
- From Electron Microscopy

Protein Structure Determination

- All structures available are ‘models’.
- A model is simply a hypothesis based on the available data.

- THE MORE DATA YOU HAVE, THE MORE RELIABLE THE MODEL

X-ray crystallography

For *small molecules* (not macromolecules)

- At 0.8 Angstrom resolution
  - 10 observations/parameter (9 parameters)
  - i.e. **90 'measurements'** per atom.
- At 1.2 Angstrom resolution
  - 5 observations/parameter (4 parameters)
  - i.e. **20 'measurements'** per atom.

X-ray crystallography

For proteins …

- At 2.0 Angstroms resolution
  - 2 observations/parameter (4 parameters)
  - i.e. **8 measurements per atom**

Number of observations is improved by including **geometry restraints** from knowledge of protein structure.

NMR Spectroscopy

- Good small molecule X-ray structure has 90 measurements per atom.
- Reasonable protein X-ray structure has 20 measurements per atom.
- **Good NMR protein structure has 10 measurements *per residue.***

NMR Spectroscopy *versus* X-ray

- NMR has fewer measurements per residue than X-ray crystallography on which to base the model.
- NMR techniques rely more heavily on basic geometry of amino acids and protein backbone.

So, what about structure prediction?

Methods for protein structure determination

Method	**Measurements per residue**
X-Ray crystallography at 2.0 Angstroms	80 + geometry
Good NMR structure	10 + geometry
***Ab-initio* prediction**	**0 + geometry!!**

Protein structure prediction

Since there are 0-few measurements, the challenge is to find constraints on the structure from somewhere...

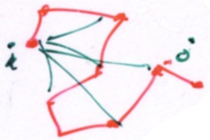
- Homology to known structure.
- Distance restraints (disulphides etc).
- Restraints from secondary structure prediction.
- Buried/Exposed restraints.

Ab initio structure prediction

Ab initio prediction - energy calculations

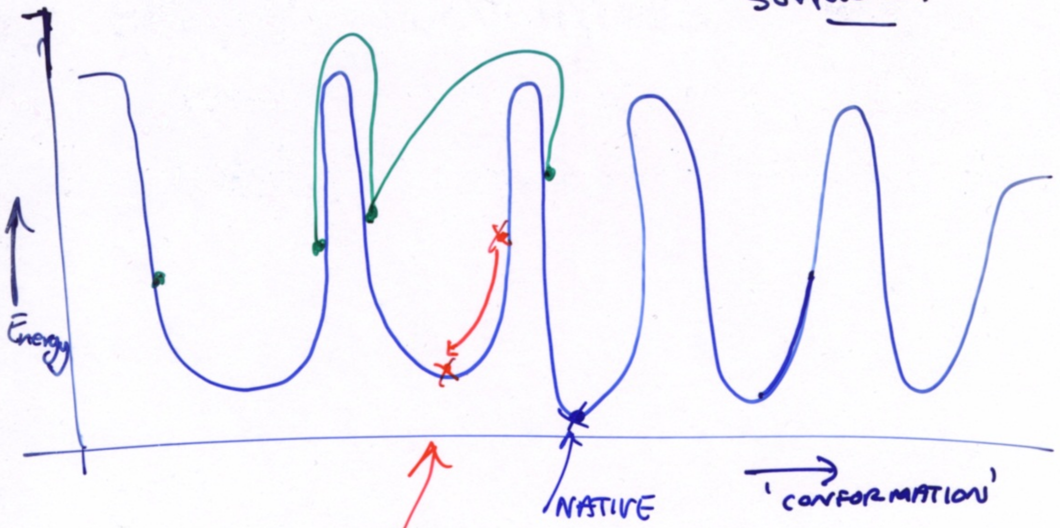
- Attempt to ‘fold’ protein from extended chain.
- Model all inter-atomic interactions as an energy term.
- Find the conformation that minimises the sum of pair-wise interaction energies.

ENERGY CALCULATIONS



UNFORTUNATELY

— EGG BOX Potential energy surface.



→ ENERGY MINIMISATION / MOLECULAR DYNAMICS

S.A. - LARGE RADIUS OF CONVERGENCE

Ab initio - energy calculations Problems

- Potentials.
- Search space is large.
- Multiple minima - distinguishing native from non-native conformations

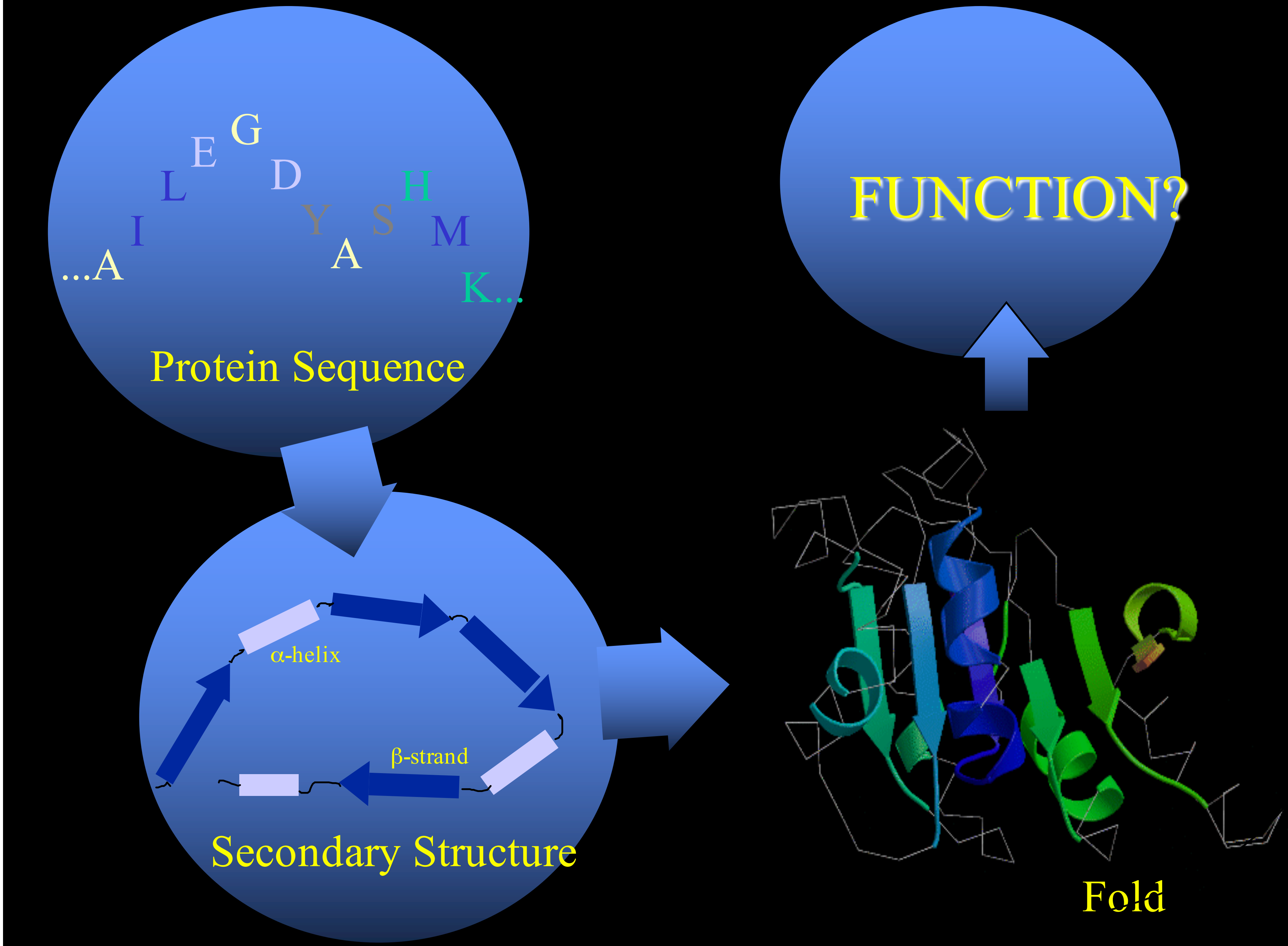
Ab initio - energy calculations

- So far only limited success in conjunction with strong restraints. For example:
  - Coiled coils.
  - Disulphide rich proteins.

Hybrid *Ab initio* methods

- Fragment assembly – David Baker – Rosetta and "folding at home" – see Baker Lab web site.

- Generate large number of small fragments of structure consistent with a simplified potential and known folding units
- Cluster millions of generated structures to find most consistent folds.
- Also see *Nature* 8th Nov 2007 for application in X-ray crystallography.



Protein Sequence

Secondary Structure

Fold

FUNCTION?

Protein Structure Prediction *Homology modelling*

- **Exploits principle that “similarity of sequence implies similarity of 3D structure”**
- Relies on finding similarity to protein of known structure.
- Must get alignment between known and unknown correct.
- Build loops correctly.

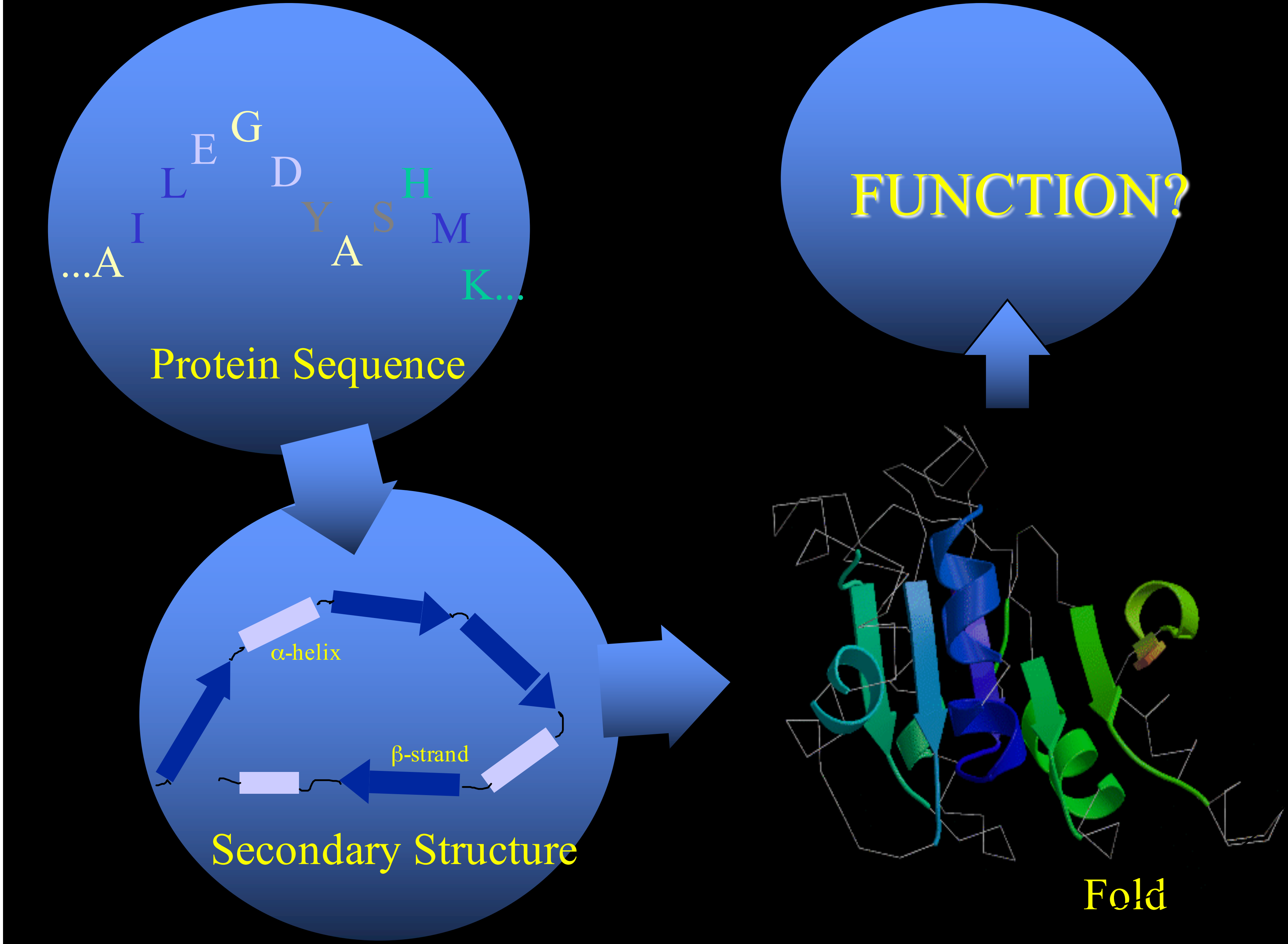
Resources for Homology Modelling

- Swissmodel
  - http://swissmodel.expasy.org/
- ModBase
  - http://modbase.compbio.ucsf.edu
- PhyreII
  - Imperial College – Google it!

Homology Modelling

1. Find homologue of known structure (scaffold)
2. Accurately align target sequence to scaffold protein sequence
3. Substitute amino acids in conserved core of scaffold protein for those in the target
4. Adjust side-chain conformations to remove clashes and optimise geometry
5. Model loops in the target that are of different length to those in the scaffold
6. Build multiple models where there is uncertainty over any of the above stages
7. Check model for gross errors (buried charges/exposed hydrophobics etc)

Protein Secondary Structure Prediction



Protein Sequence

Secondary Structure

Fold

FUNCTION?

Protein Secondary Structure Prediction

Useful, even if not used to gain a full three-dimensional structure.

Identify ‘safe’ regions for mutagenesis

Help improve sequence searches/fold recognition.

Improve sequence alignment accuracy.

A Few Ideas That Have Been Tried in Protein Secondary Structure Prediction

- Sequence Similarity: Do a sequence alignment against a protein of known structure.
- By eye: Look for patterns of conservation in a multiple sequence alignment.
- Statistical Methods: Singlets (Chou & Fasman), Information theory:(GORI,II,II,IV...) Bayes etc. etc. etc....
- Pattern Matching: Templates (Taylor), AI (Cohen) etc. ...
- Neural Nets: Karplus, Rost & Sander, etc...
- Nearest Neighbour: Salamov & Solovyev, Levin, etc ...
- Others: Tarot, Tea Leaves etc ...

A Few Ideas That Have Been Tried in Protein Secondary Structure Prediction

- Sequence Similarity: Do a sequence alignment against a protein of known structure.
- By eye: Look for patterns of conservation in a multiple sequence alignment.
- Statistical Methods: Singlets (Chou & Fasman),
  Information theory:(GORI,II,II,IV...)
  Bayes etc. etc. etc....
- Pattern Matching: Templates (Taylor), AI (Cohen) etc. ...
- Neural Nets: Karplus, Rost & Sander, etc...
- Nearest Neighbour: Salamov & Solovyev, Levin, etc ...
- Others: Tarot, Tea Leaves etc ...

A Few Ideas That Have Been Tried in Protein Secondary Structure Prediction

- Sequence Similarity: Do a sequence alignment against a protein of known structure.
- By eye: Look for patterns of conservation in a multiple sequence alignment.
- Statistical Methods: Singlets (Chou & Fasman), Information theory:(GORI,II,II,IV...) Bayes etc. etc. etc....
- Pattern Matching: Templates (Taylor), AI (Cohen) etc. ...
- Neural Nets: Karplus, Rost & Sander, etc...
- Nearest Neighbour: Salamov & Solovyev, Levin, etc ...
- Others: Tarot, Tea Leaves etc ...

If the protein is similar to a protein of known three dimensional structure, then just use alignment…

Prediction by alignment to protein of known structure.

No poor predictions by alignment for similarity > 6 sigma.

P.E.Boscott *et al.*

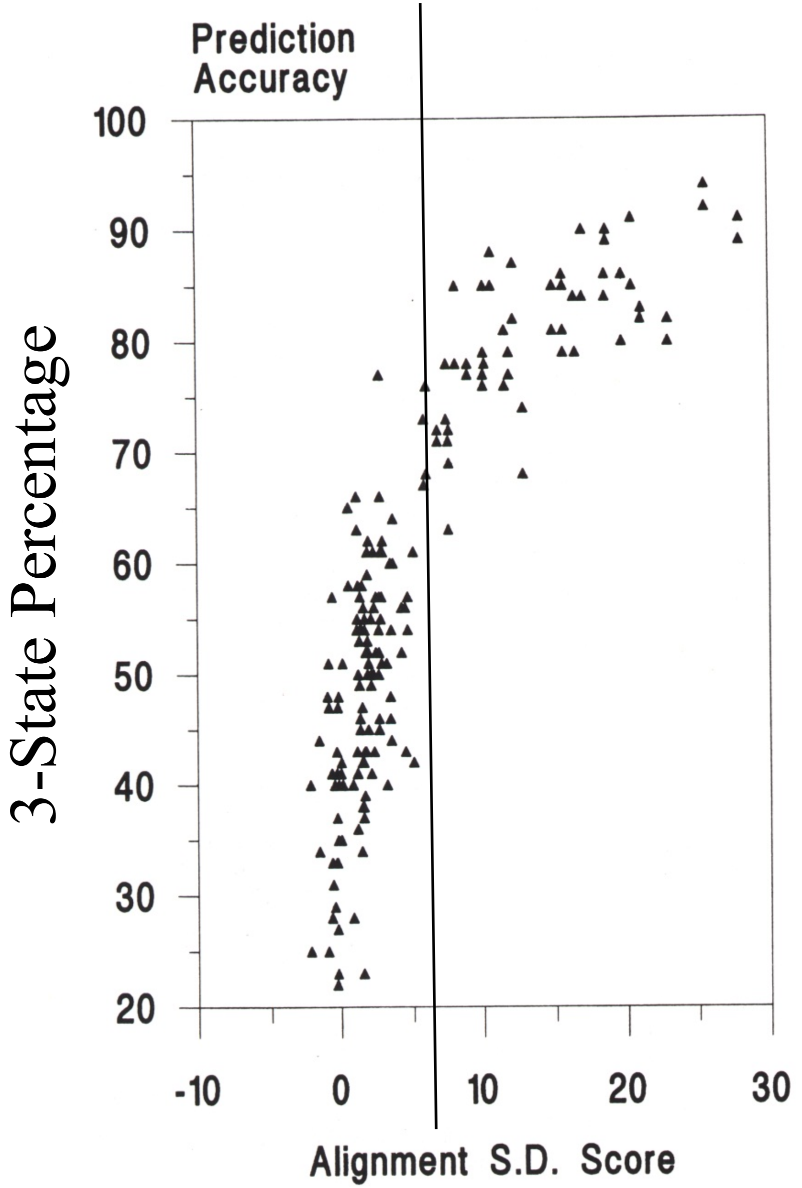


Fig. 2. The accuracy of secondary structure prediction by sequence alignment plotted against the alignment SD score to the homologous protein. One hundred and eighty-two predictions were made from pairwise alignment of the proteins in Table I.

Boscott, P. E., Barton, G. J. and Richards, W. G. (1993), *Prot. Eng.,* **6**, 261-266.

Prediction methods normally need

- To be trained on some data
- Tested on different data

For secondary structure prediction:

- Train on a subset of the Protein Data Bank (PDB) of known protein three-dimensional structures

- Test on different subset of the PDB that was not used in training

- Best test is on proteins for which no structure is known.
  - However, you have to wait for someone to solve the structure by X-ray methods before you can find out the result

A Few Ideas That Have Been Tried in Protein Secondary Structure Prediction

- Sequence Similarity: Do a sequence alignment against a protein of known structure.
- By eye: Look for patterns of conservation in a multiple sequence alignment.
- Statistical Methods: Singlets (Chou & Fasman), Information theory:(GORI,II,II,IV...) Bayes etc. etc. etc....
- Pattern Matching: Templates (Taylor), AI (Cohen) etc. ...
- Neural Nets: Karplus, Rost & Sander, etc...
- Nearest Neighbour: Salamov & Solovyev, Levin, etc ...
- Others: Tarot, Tea Leaves etc ...

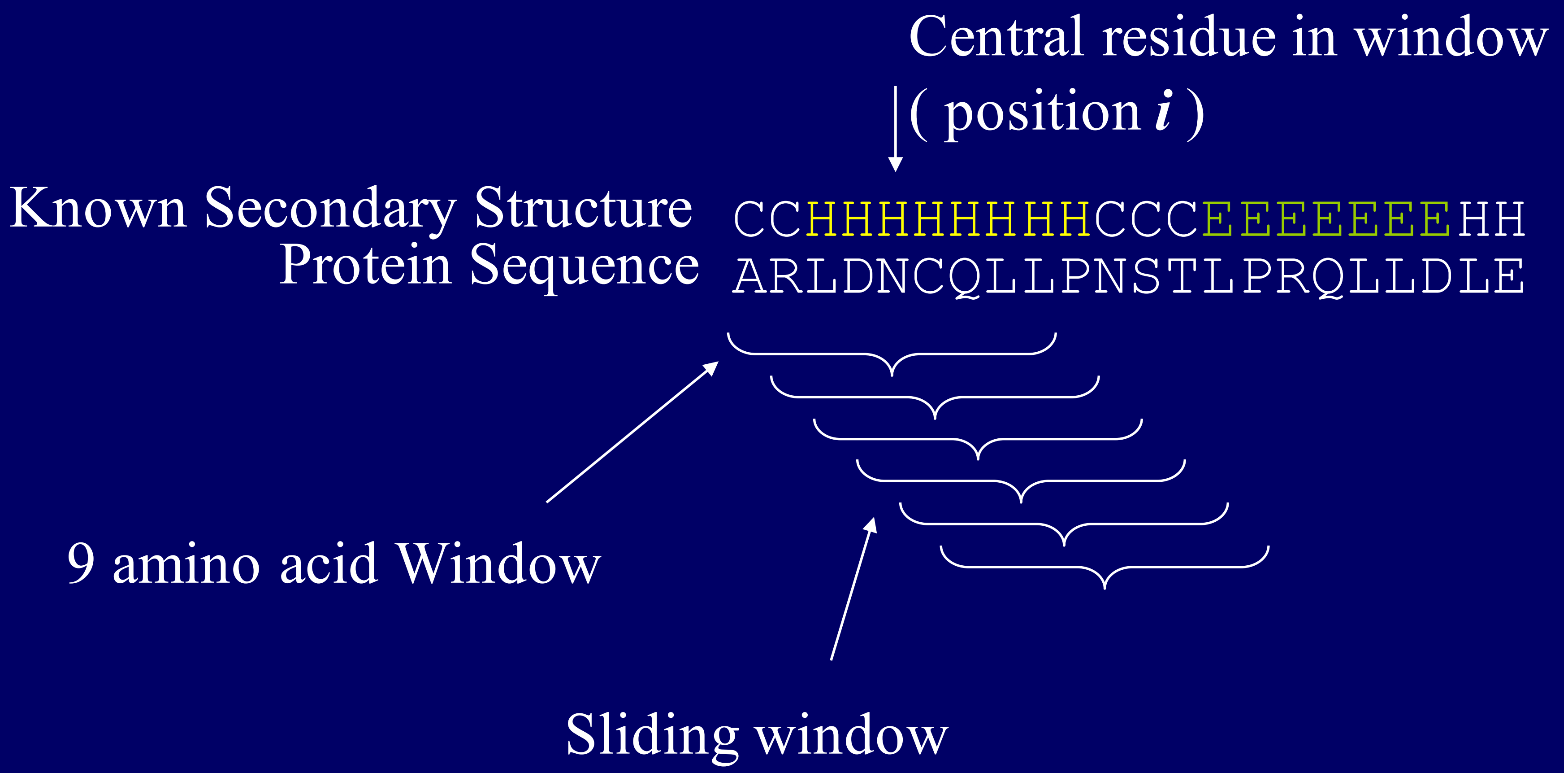
Statistical Methods

- Simplest method - singlets:
  - Look at proteins of known 3D structure
  - Count how many times each amino acid is seen in each secondary structure state. For example, Ala in Helix, Strand, Turn etc
  - Express as a probability that each amino acid will appear in that state

Statistical Methods

- More complex method - windowing:
  - Take a sliding window of say, 9 amino acids
  - Gather statistics on all 9 positions relative to the secondary structure state of the central position
  - Express as a probability or “information content”

More complex method - windowing



Central residue in window

(position ***i***)

Known Secondary Structure CCHHHHHHHHCCCEEEEEEEHH

Protein Sequence ARLDNCQLLPNSTLPRQLLDLE

9 amino acid Window

Sliding window

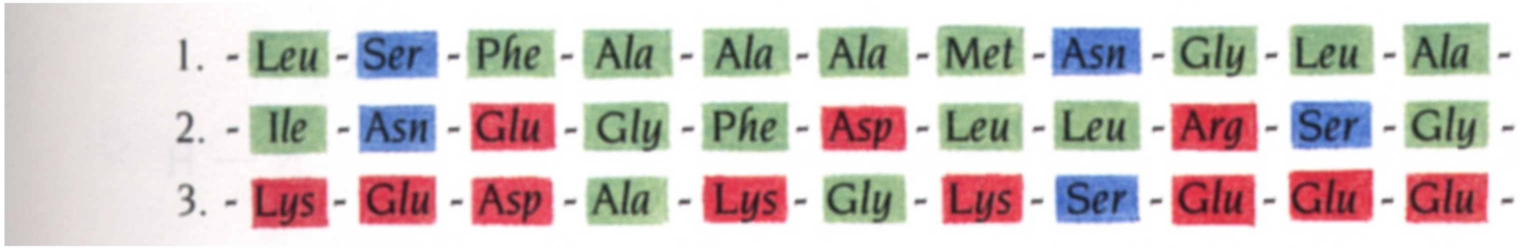
Gather statistics about how often each amino acid is seen at positions

i-4, i-3, i-2, i-1, i, i+1, i+2, i+3, i+4

For each secondary structure type at position *i*.

A Few Ideas That Have Been Tried in Protein Secondary Structure Prediction

- Sequence Similarity: Do a sequence alignment against a protein of known structure.
- By eye: Look for patterns of conservation in a multiple sequence alignment.
- Statistical Methods: Singlets (Chou & Fasman), Information theory:(GORI,II,II,IV…) Bayes etc. etc. etc.…
- Pattern Matching: Templates (Taylor), AI (Cohen) etc. …
- Neural Nets: Karplus, Rost & Sander, etc...
- Nearest Neighbour: Salamov & Solovyev, Levin, etc …
- Others: Tarot, Tea Leaves etc …



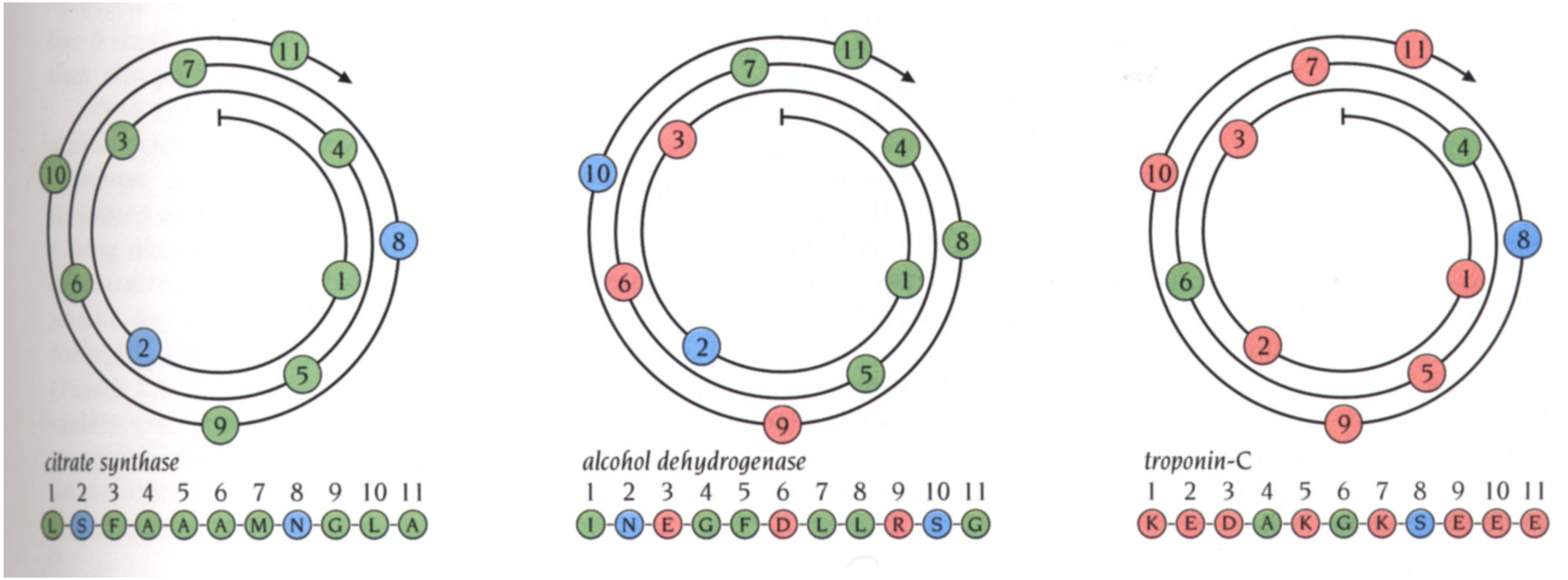
1. - Leu - Ser - Phe - Ala - Ala - Ala - Met - Asn - Gly - Leu - Ala -

2. - Ile - Asn - Glu - Gly - Phe - Asp - Leu - Leu - Arg - Ser - Gly -

3. - Lys - Glu - Asp - Ala - Lys - Gly - Lys - Ser - Glu - Glu - Glu -

1. Buried helix; 2. part exposed helix; 3. exposed helix

Helical wheel plots to show location of hydrophobic amino acids on face of helix.



The **more data** you have, the better the prediction

- Secondary structure predictions from *multiple sequence alignments* are more accurate than from single sequences.

  Some numbers later...

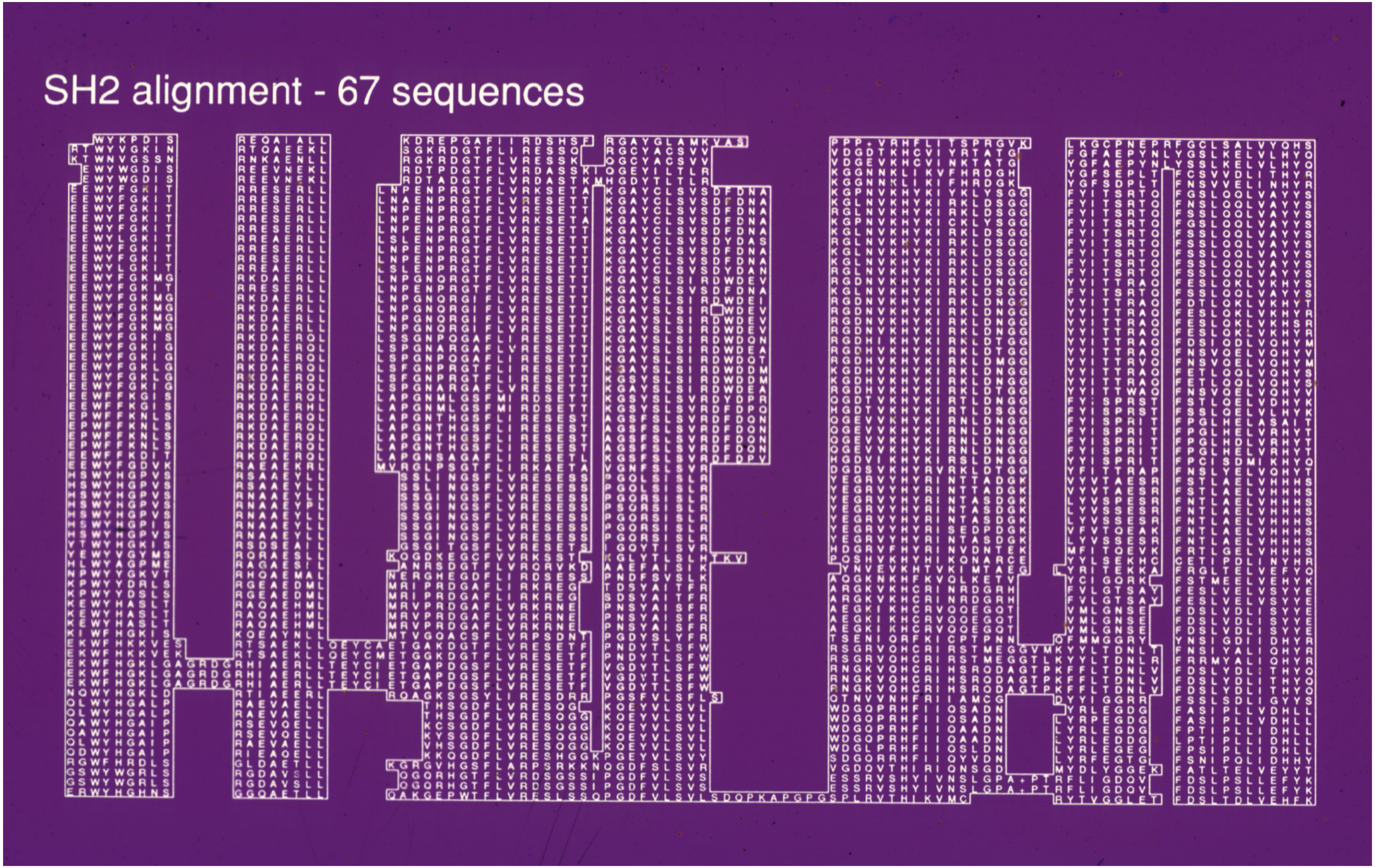
- For now, some example blind predictions.

Why predict by eye when we have programs to do this?

- Predicting by eye, helps to understand the evidence upon which an automatic prediction is based.
- Examples, from early work.

Example Prediction

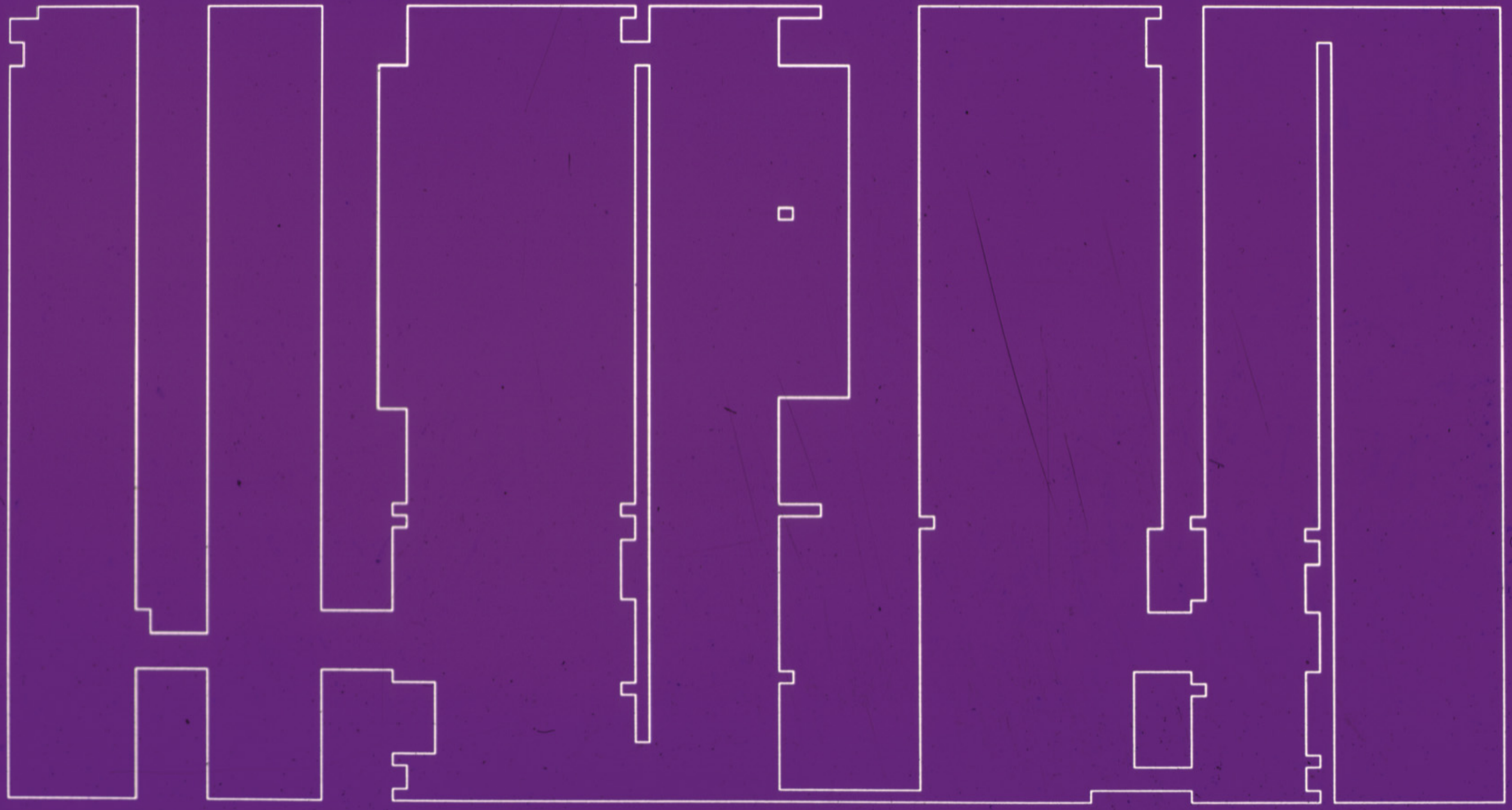
- SH2 Domains – phosphotyrosine binding domains seen in many signalling proteins



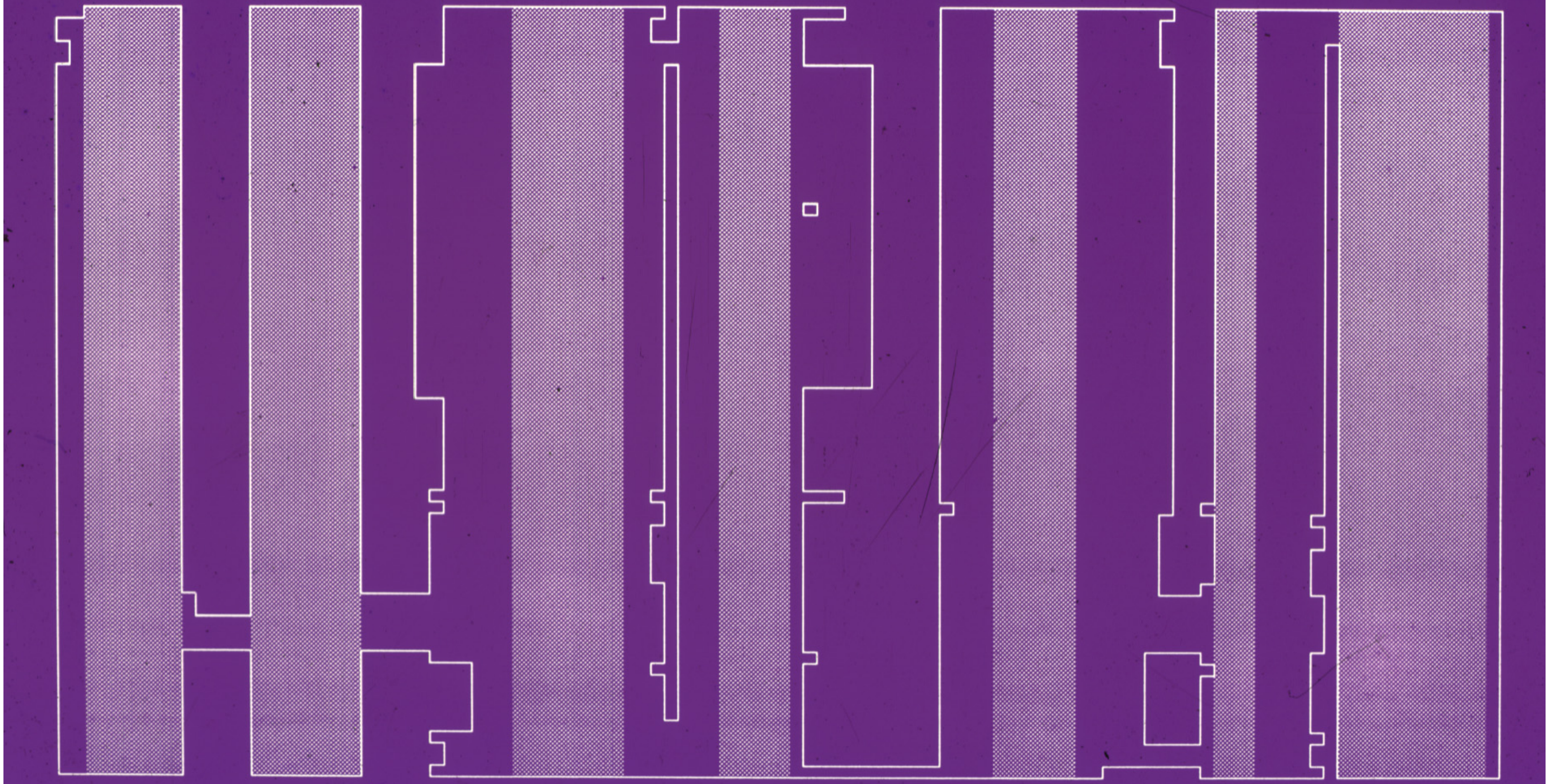
SH2 alignment - 67 sequences

Russell, R. B. Breed, J. and Barton, G. J. (1992), *FEBS Lett.*, **304**, 15-20.

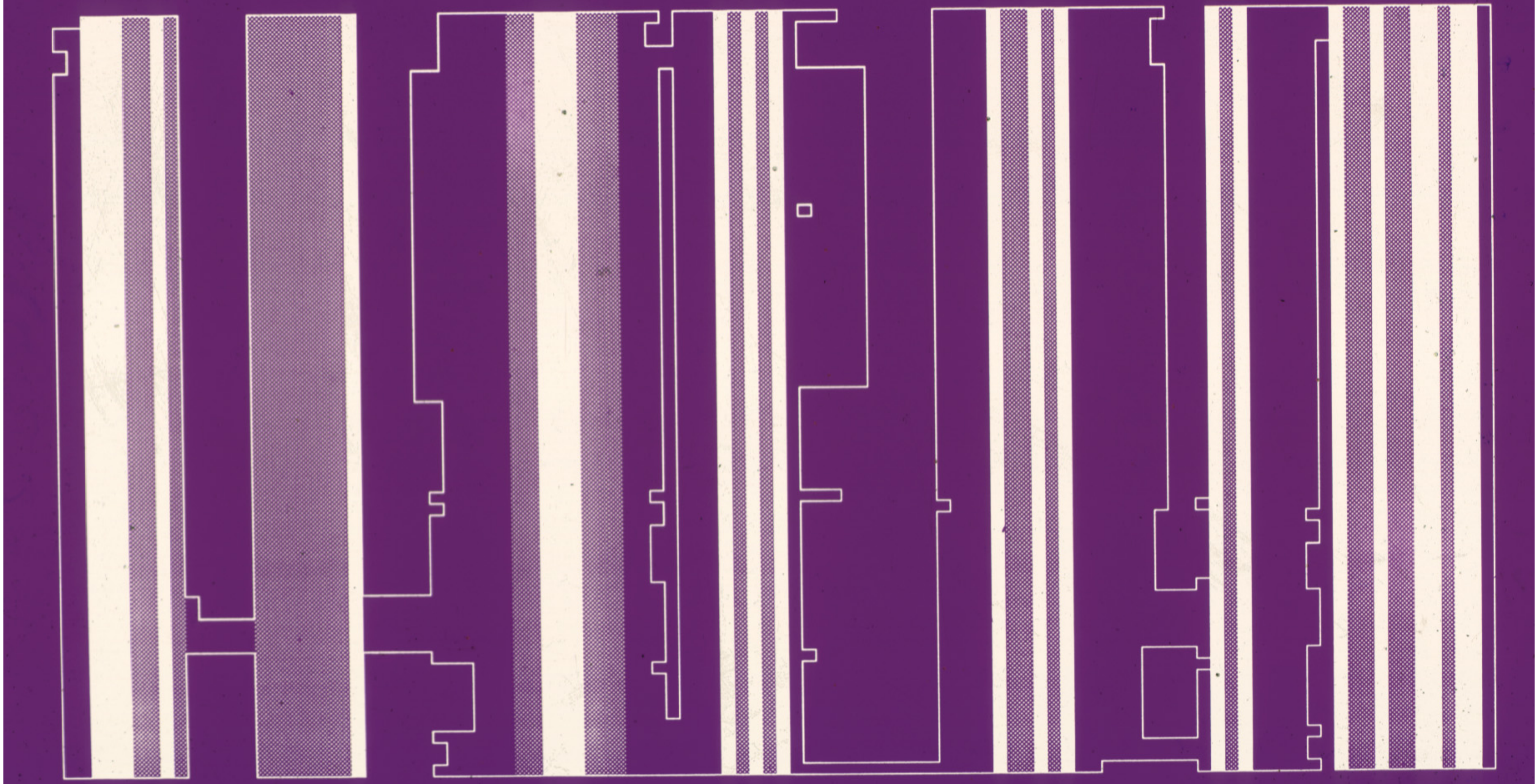
SH2 alignment - without the sequences

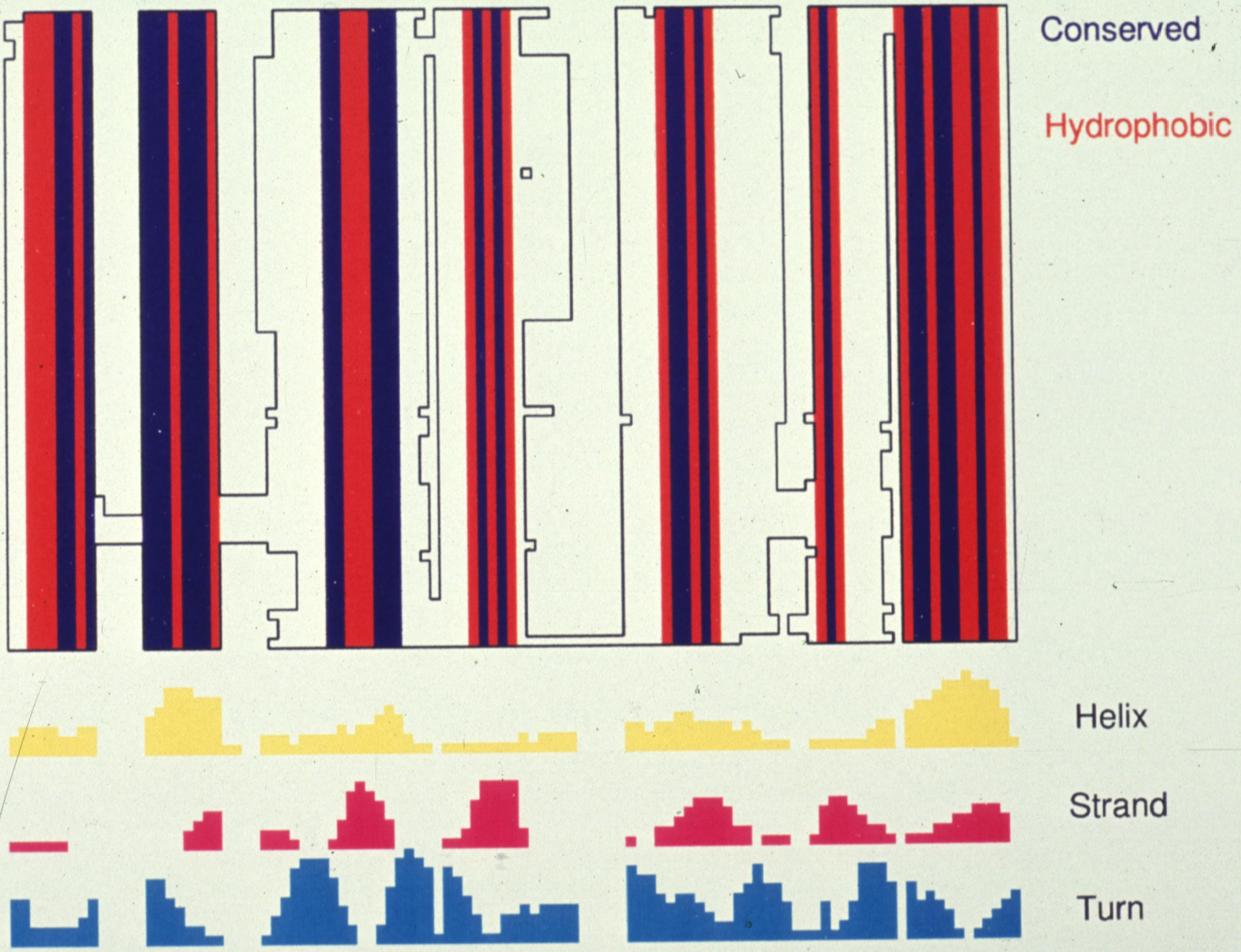


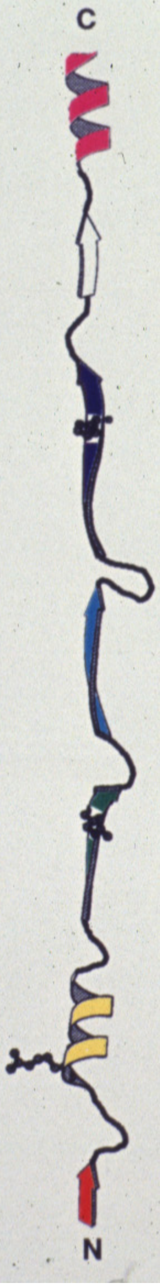
SH2 alignment - conserved regions



SH2 alignment - conserved regions and hydrophobics




Conserved
Hydrophobic
Helix
Strand
Turn



1, 4, 5, 8 conservation - surface helix

Short Strand possibly buried

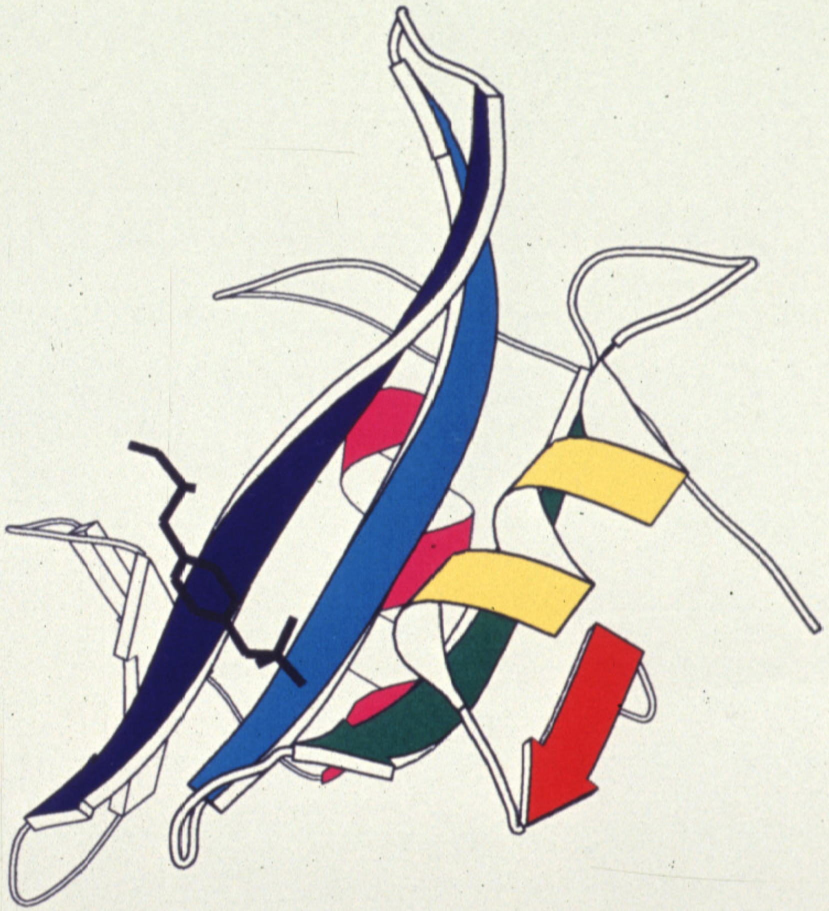
Alternating hydrophobics suggest surface strand

Alternating hydrophobics suggest surface strand

Buried strand with exposed face at end

No clear hydrophobic pattern

Run of 3 conserved hydrophobics suggest strand



src SH2-Domain

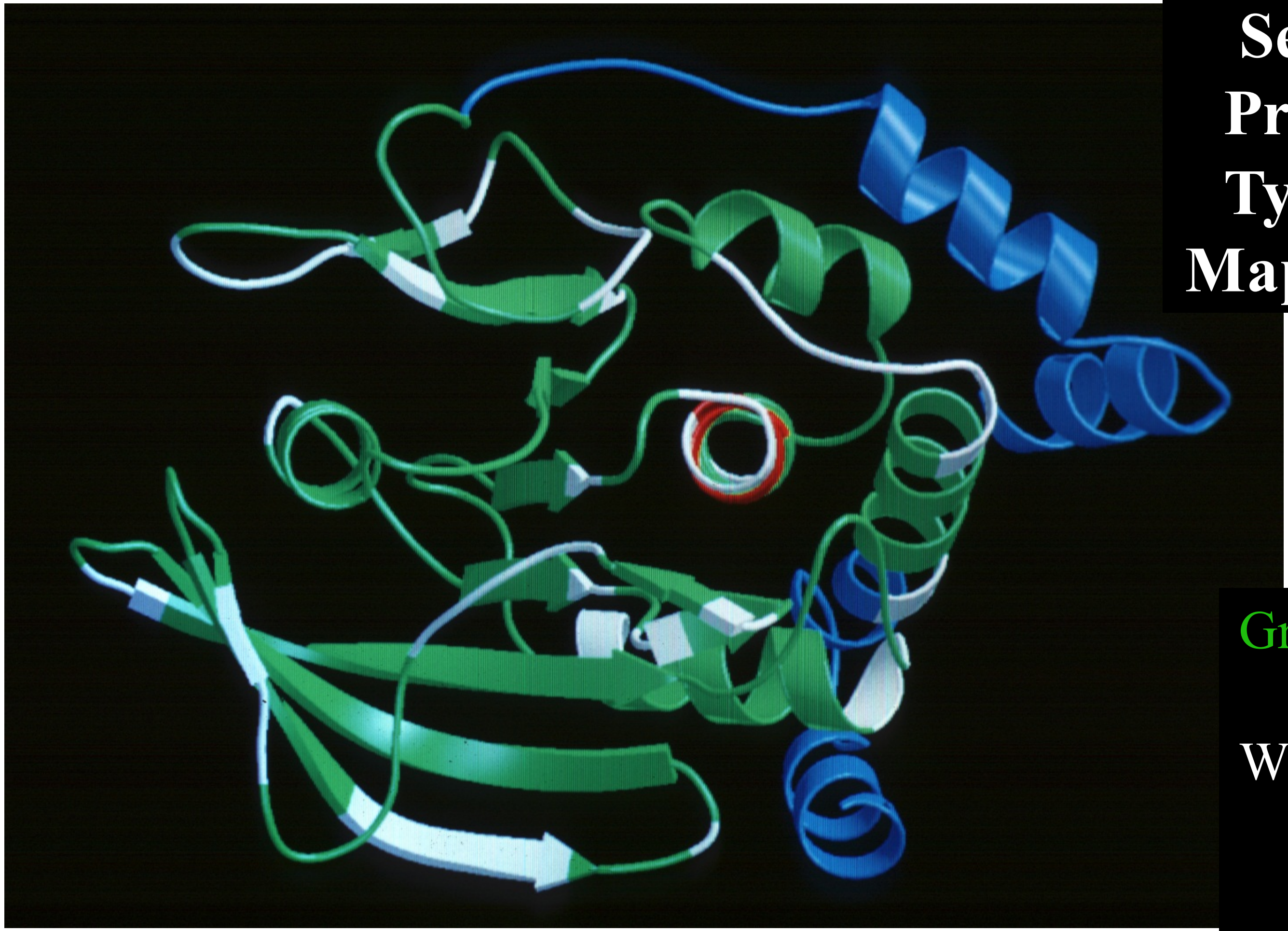
(Waksman et al, Nature, 358, 646-653, 1992)

Simple rules for secondary structure prediction by eye

Given an *accurate* multiple sequence alignment

- Positions of insertions and deletions - coil.
- Conserved Gly/Pro - turn.
- Run of conserved hydrophobics - buried β.
- Alternating conserved residues - surface β.
- 1,4,5,8 pattern of conserved residues - α.
- ‘pairs’ of conserved residues - α.

Blind Secondary Structure Prediction for Protein Tyrosine Phosphatase Mapped on to Structure.



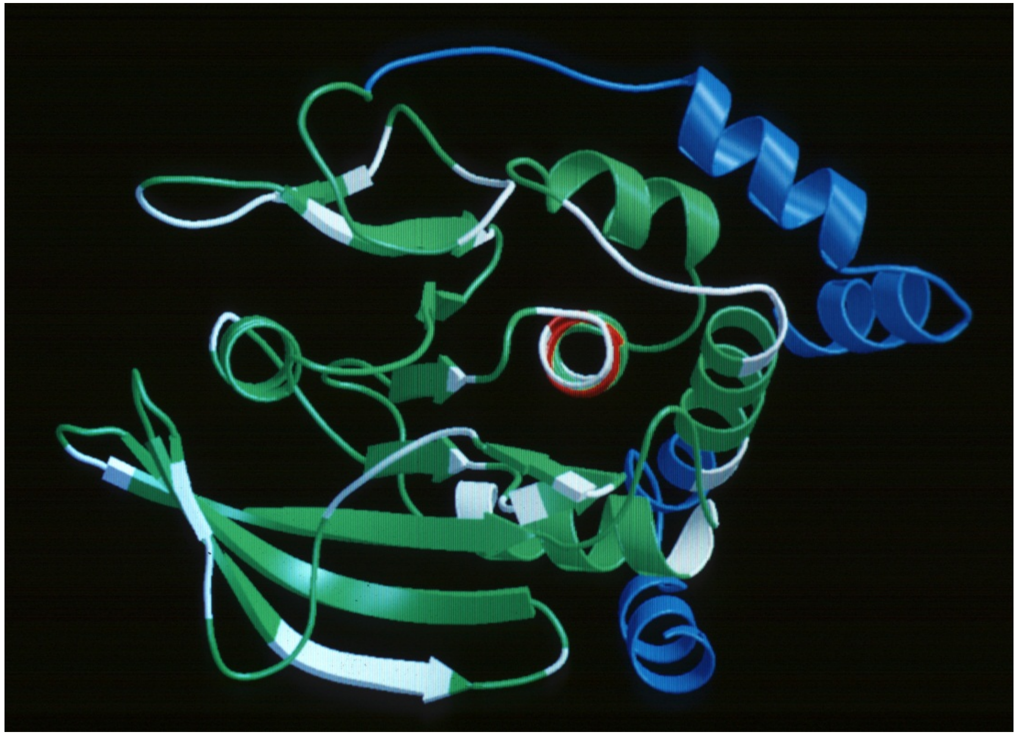
>75% Accuracy

Green: Correct Prediction.

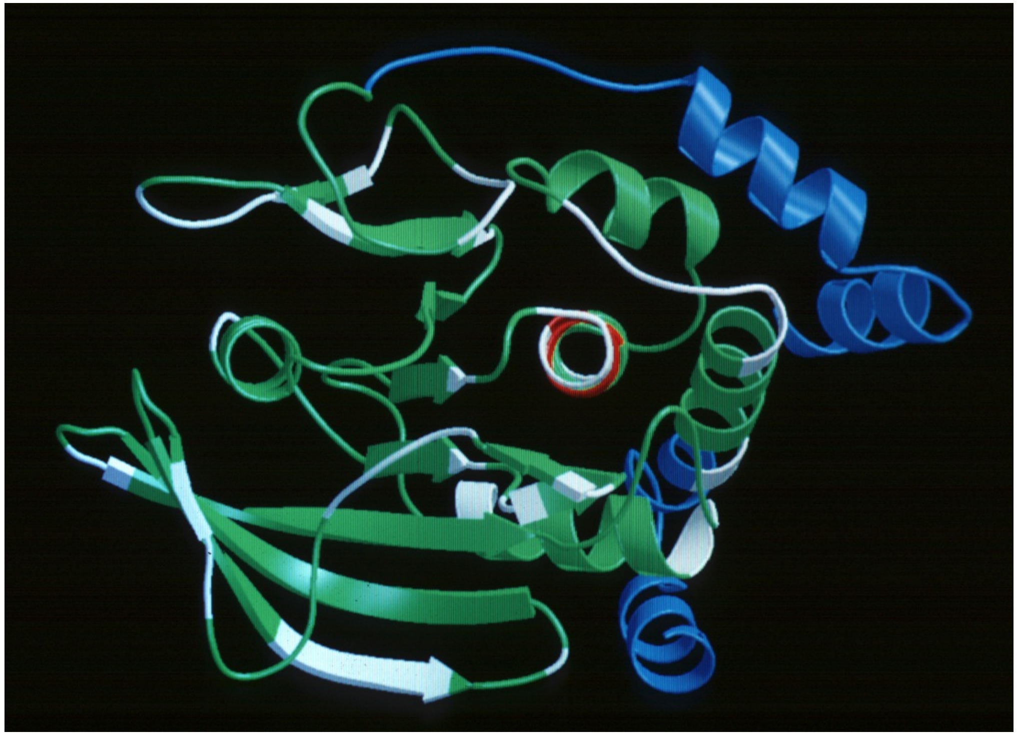
White: Coil for strand/helix or Strand/helix for coil

Red: Helix for strand or Strand for Helix

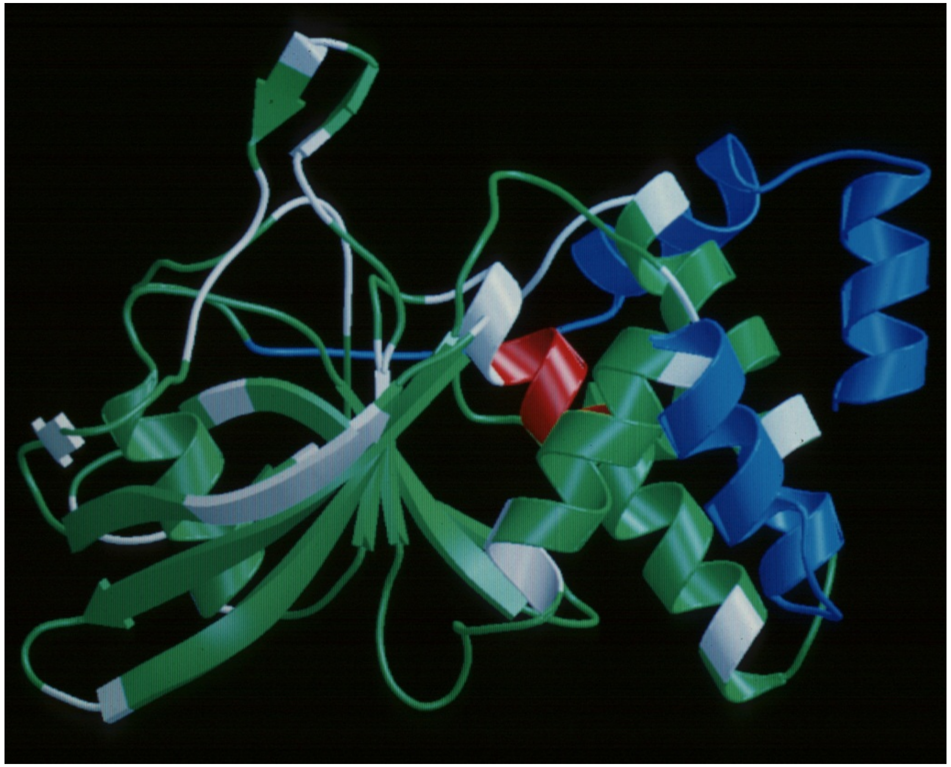
Blue: Not predicted.



Yersinia Protein Tyrosine Phosphatase Prediction
> 75% Accuracy



Yersinia Protein Tyrosine Phosphatase Prediction
> 75% Accuracy



Oh Dear!

End