Royal Holloway
University of London

The detection of DNA-binding proteins by means of structural motifs

Hugh Shanahan
Department of Computer Science
Royal Holloway, University of London

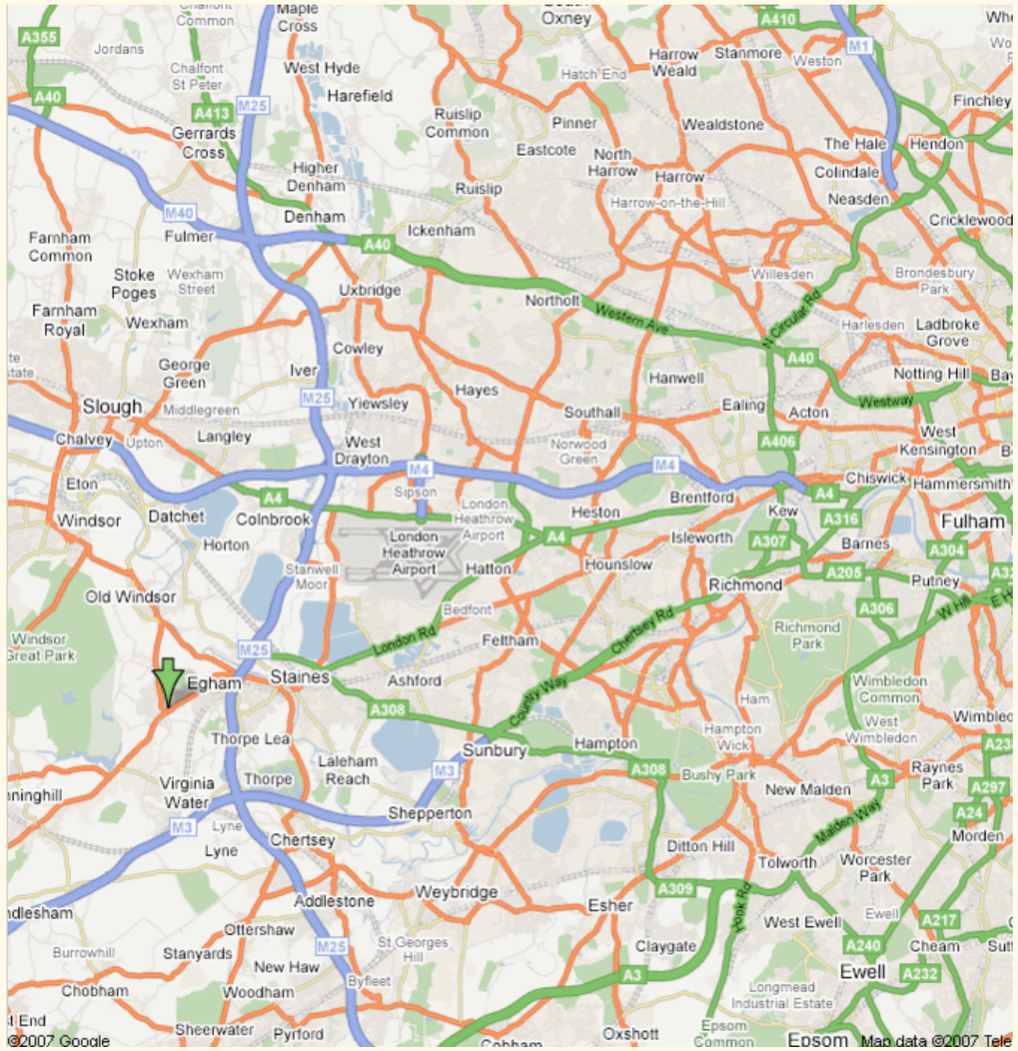
University of Glasgow
24 January 2008

A little bit about Royal Holloway ...

- Part of the University of London Federation
- In very leafy Surrey
- Most well known for its Arts Faculty
- But there’s also a Science faculty !
- CS department:
- Strong track record in Machine Learning
- Expanded Computational Biology recently

A little bit about Royal Holloway ...

- Part of the University of London Federation
- In very leafy Surrey
- Most well known for its Arts Faculty
- But there's also a Science faculty !
- CS department:
- Strong track record in Machine Learning
- Expanded Computational Biology recently



Summary

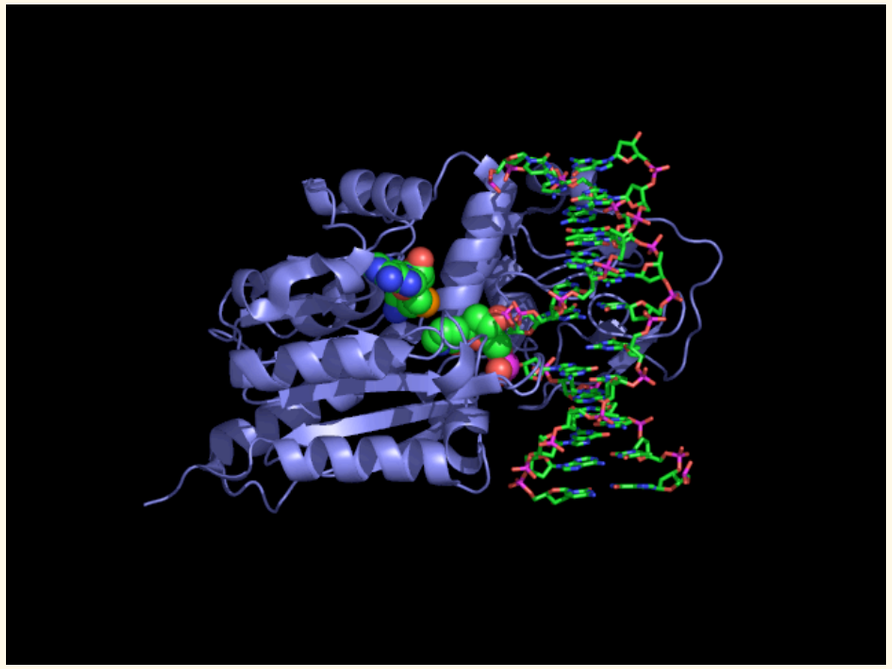
- A brief overview of DNA-binding proteins.
- Structural and Functional Genomics
- Binding Motifs: HTH, HLH, HhH
- Searching for the HTH structural motif
- Improving on structure :- electrostatics
- Lessons from convergent evolution

DNA-binding proteins

- A wide variety of different and crucial functions

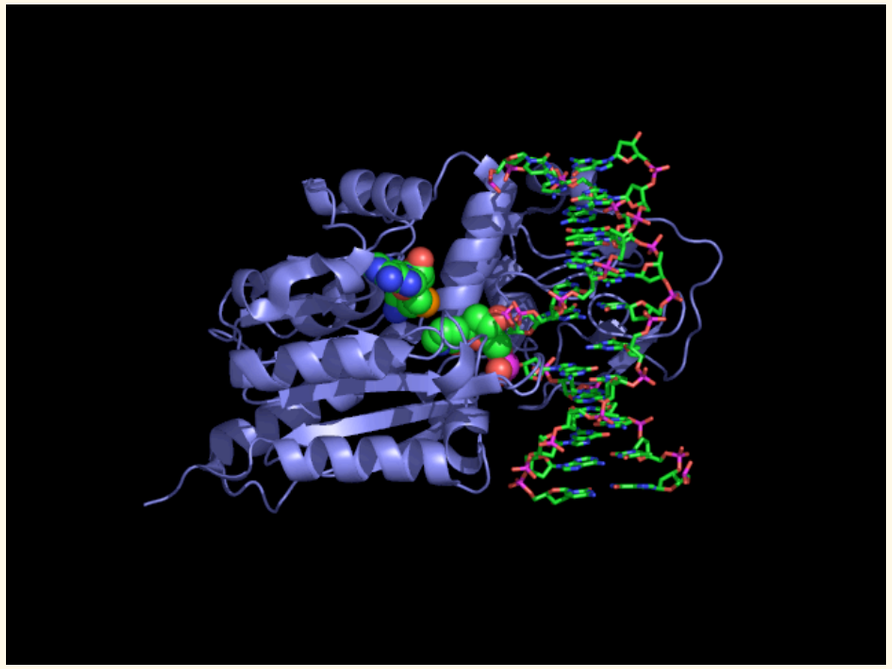
DNA-binding proteins

- A wide variety of different and crucial functions
- Enzymatic
    - Repair
    - Methylation, etc.



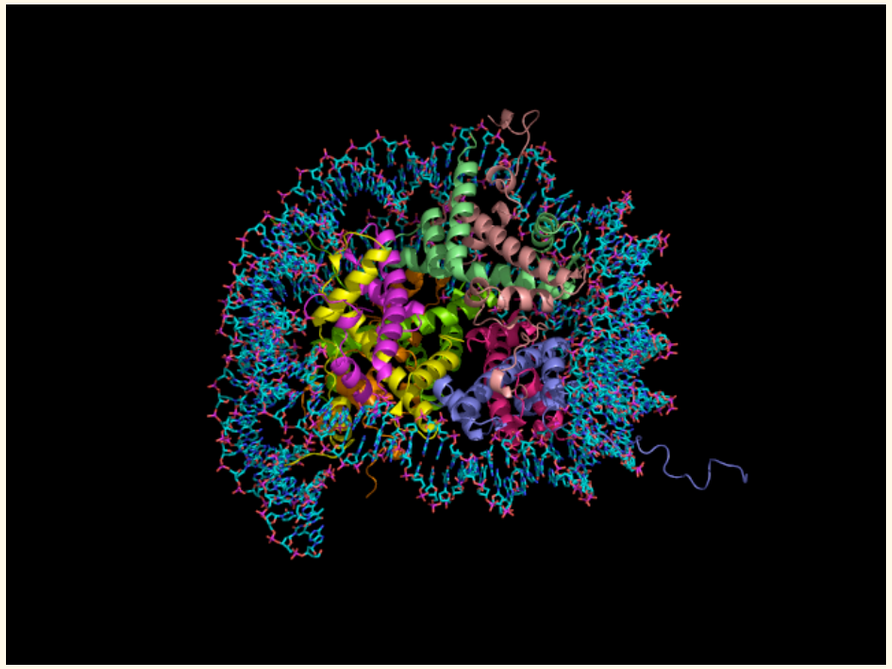
DNA-binding proteins

- A wide variety of different and crucial functions
- Enzymatic
    - Repair
    - Methylation, etc.



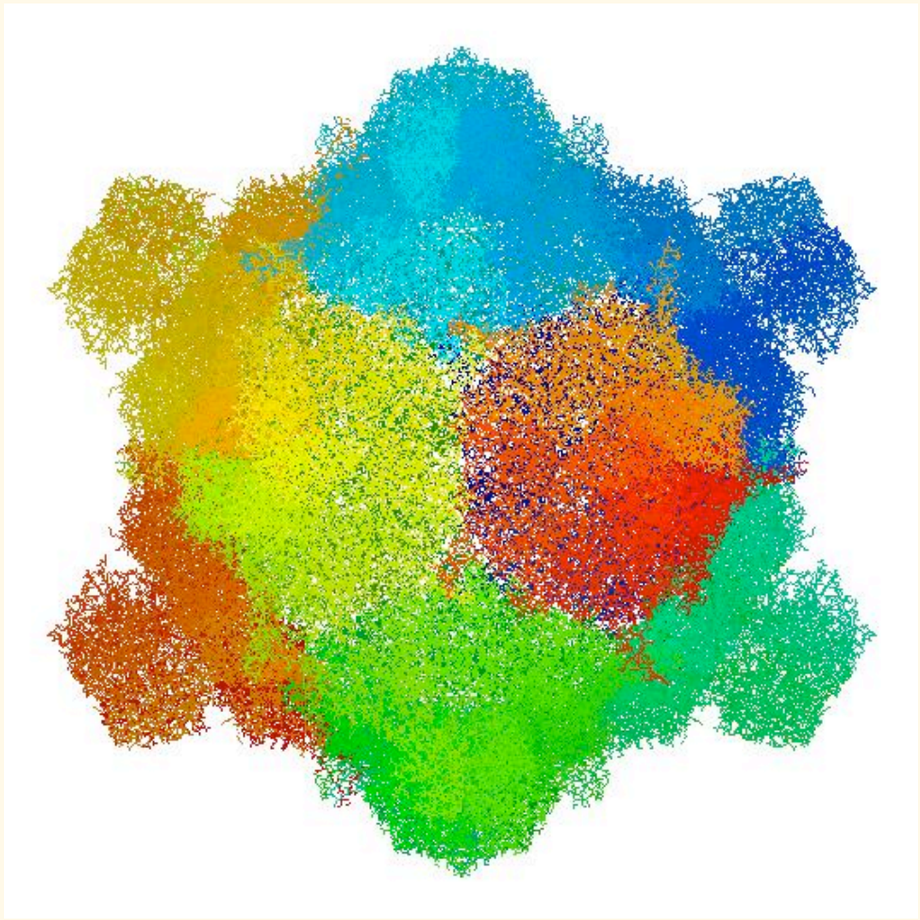
DNA-binding proteins

- A wide variety of different and crucial functions
- Enzymatic
    - Repair
    - Methylation, etc.
- Storage
    - Histones
    - Teleomeres



DNA-binding proteins

- A wide variety of different and crucial functions
- Enzymatic
    - Repair
    - Methylation, etc.
- Storage
    - Histones
    - Teleomeres
- Viral (storage)



And of course...

- Regulation of Transcription

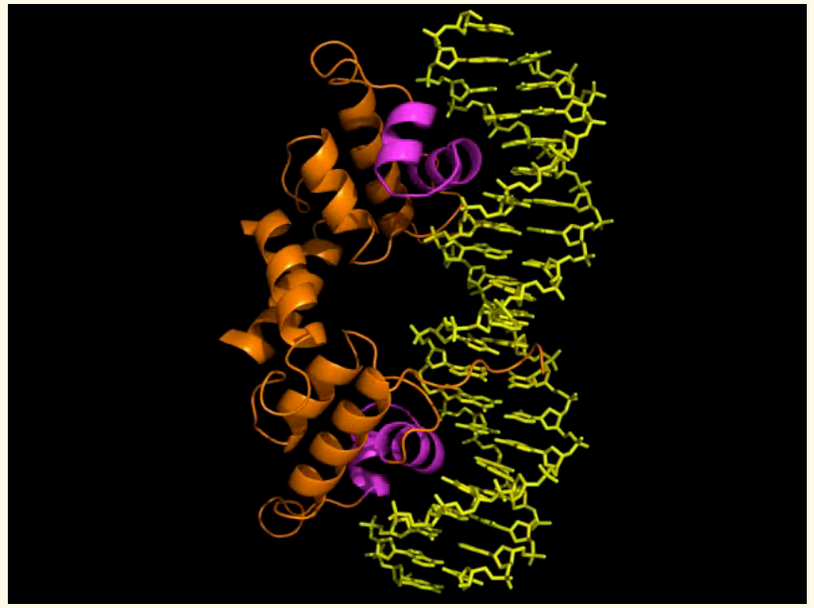
- Challenges for Systems Biology

    - Identification of consensus motifs for a given protein

    - Identification of *cis*-regulatory regions for a given gene

    - Identification of transcriptional regulatory networks

- The more DNA-binding proteins identified, the better.



Abundance

- Crude estimate:
- Eukaryotes: 6-7% of genome
- Prokaryotes: 2-3% of genome

Structural Genomics

- ‘Ab Initio’ protein structure determination remains an extremely difficult task.

- Homology modelling and modelling based on more distant homologues has made considerable progress in the last 10-15 years (c.f. CASP evaluation procedure).

- Structural Genomics: Moderately High-throughput, high accuracy determination of protein structures using X-ray crystallography and NMR.

- A goal of Structural Genomics consortia is an attempt to “fill in the gaps” in the above modelling step.

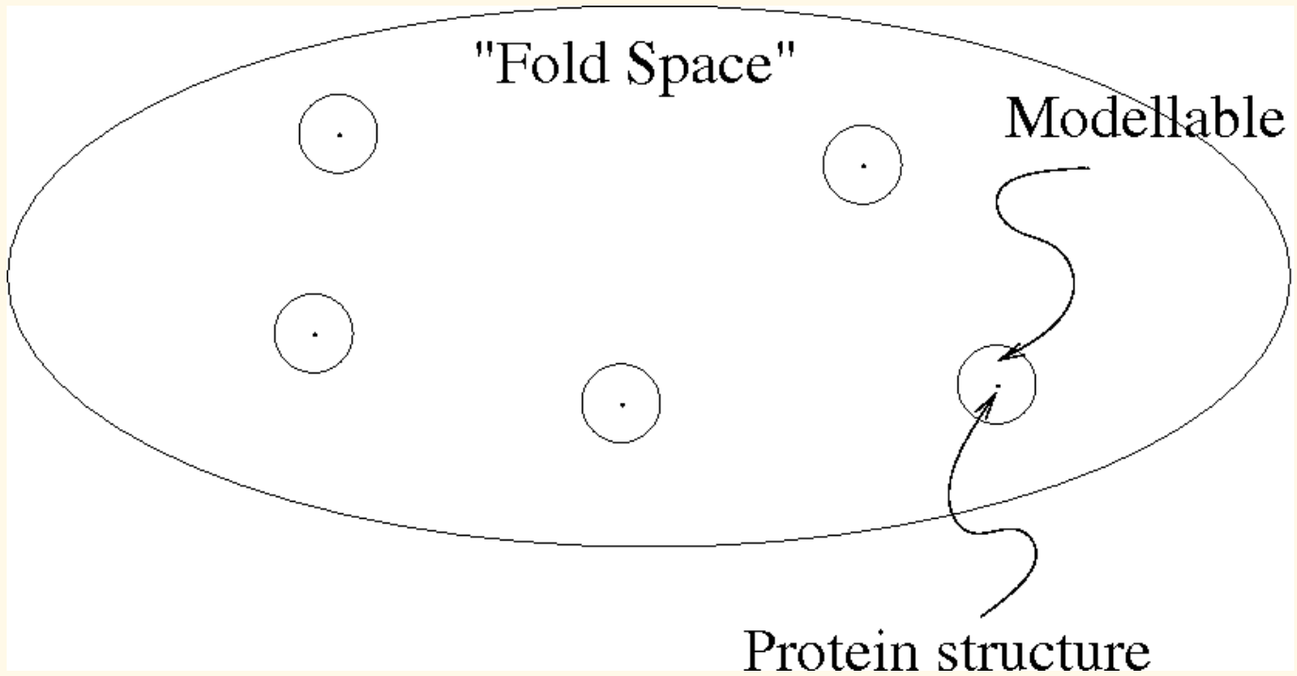
- Choice of targets: typically ones that have very low homology with known proteins.

A quandary and a opportunity

- Schematically

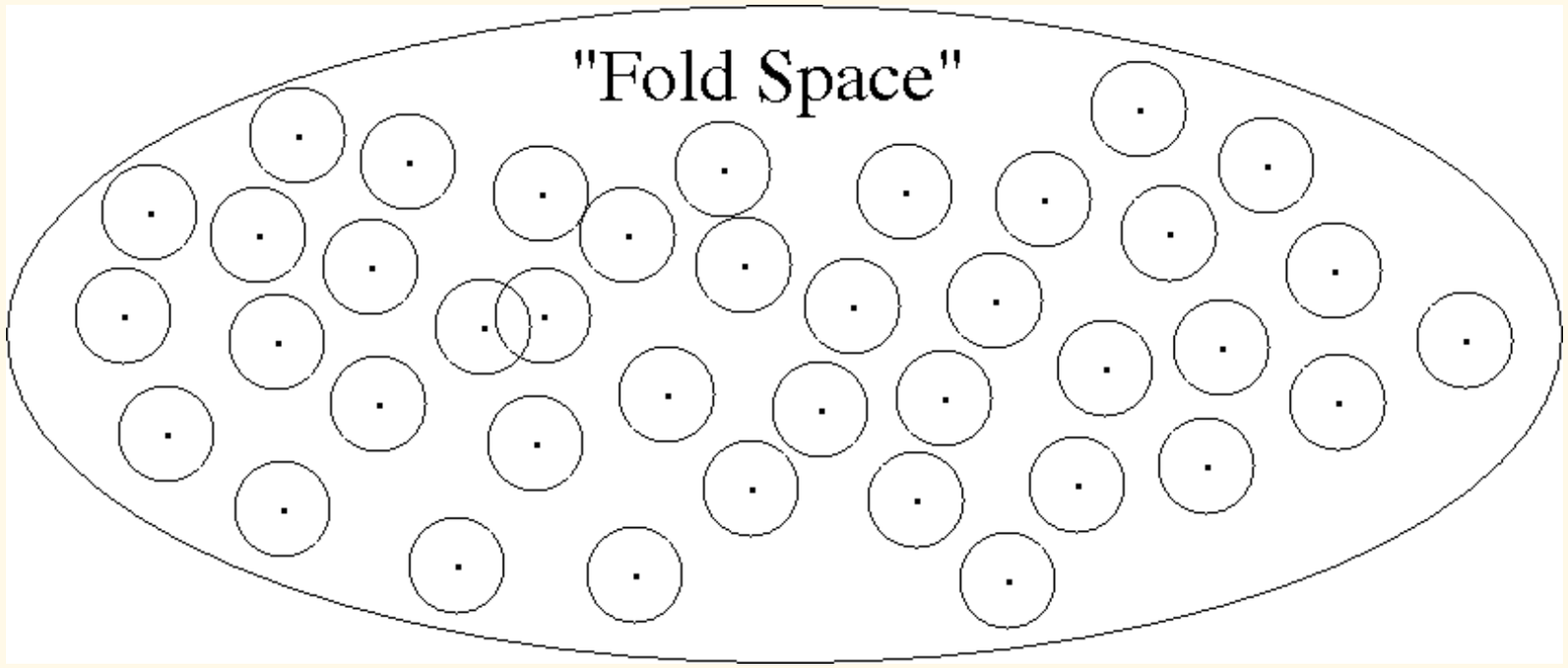
A quandary and a opportunity

- Schematically



A quandary and a opportunity

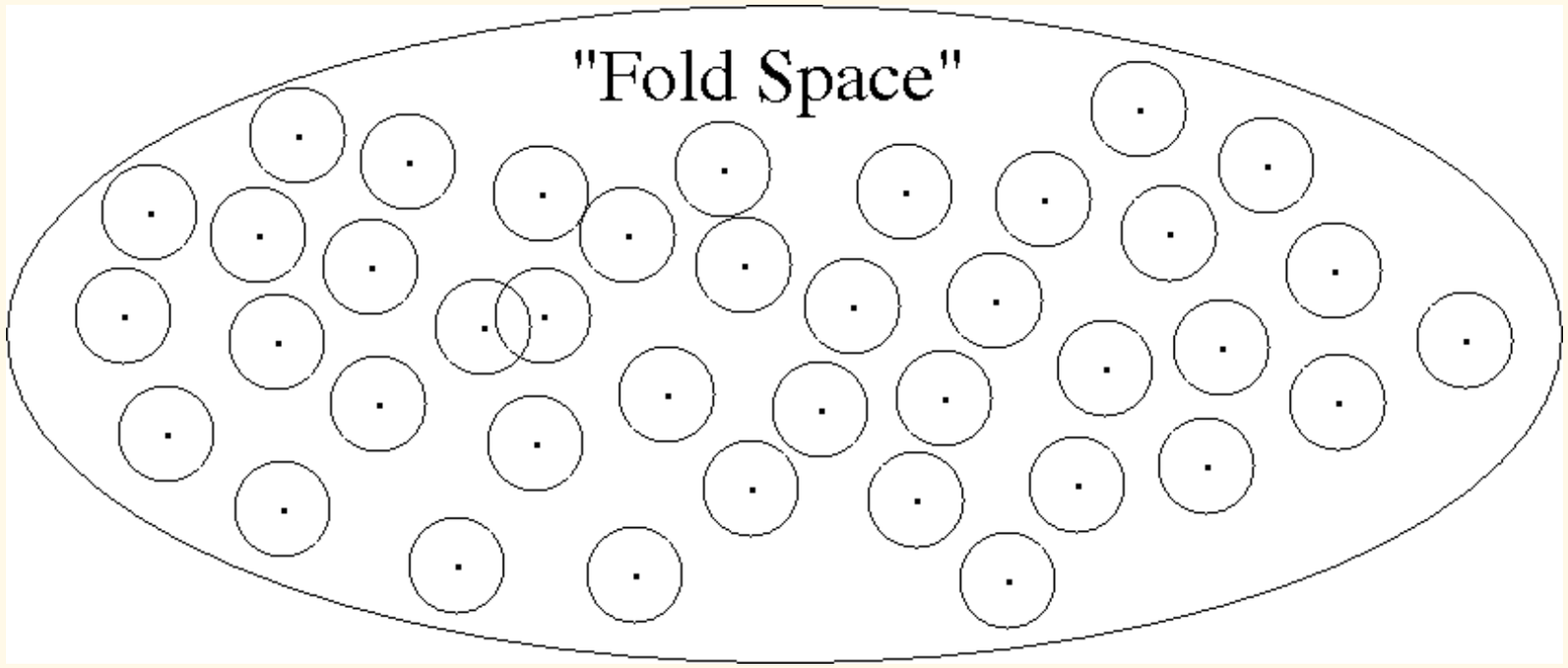
- Schematically



A quandary and a opportunity

- Schematically

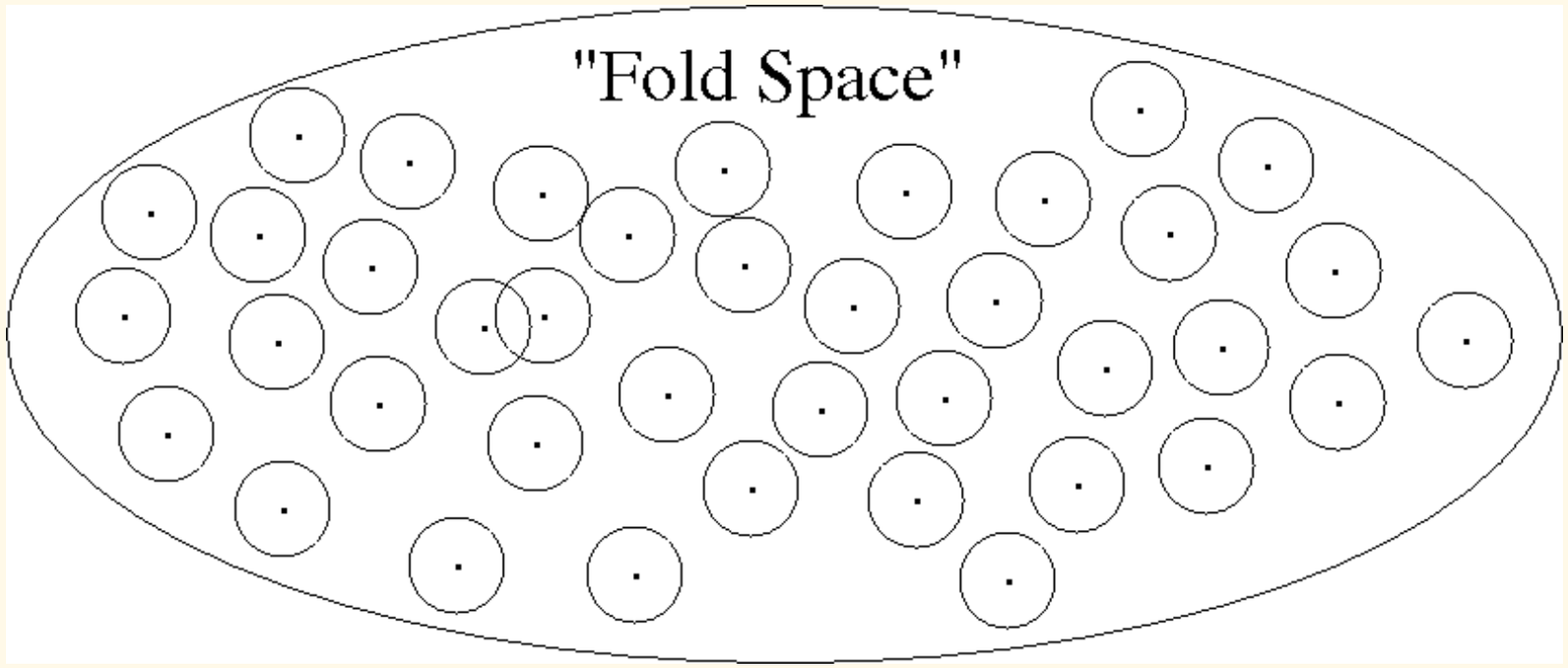
- Problem: We do not know the function of many of these targets. Typically, we cannot use homology based arguments since we won't know what the function is of any its homologues !



A quandary and a opportunity

- Schematically

- Problem: We do not know the function of many of these targets. Typically, we cannot use homology based arguments since we won't know what the function is of any its homologues !



- An opportunity: if we can determine the function (or at lest come up with a plausible short list for assay) from the structure, then we get not only this protein's function but its close homologues too !

What do we mean by “medium-throughput”

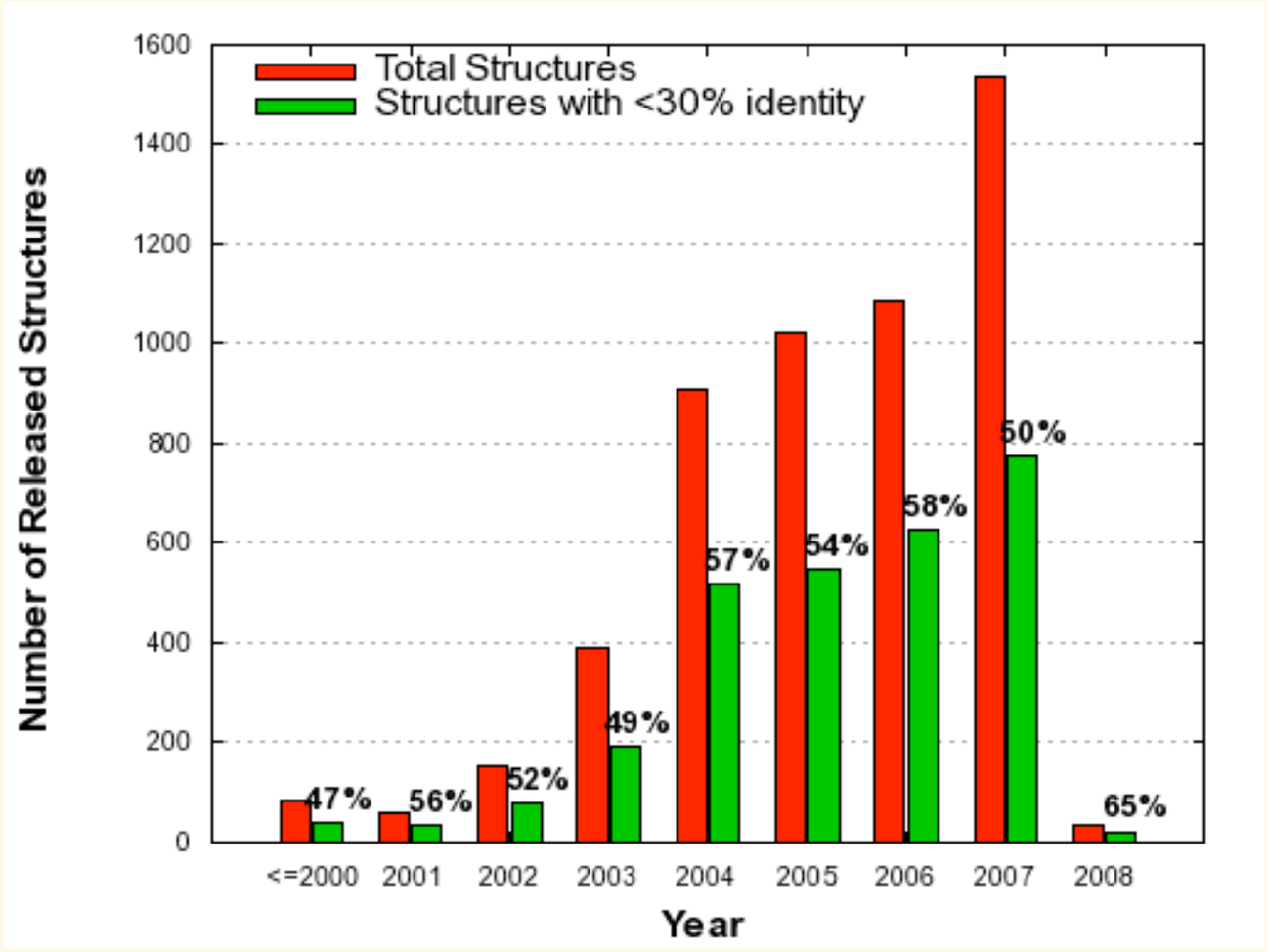
- http://targetdb.pdb.org/statistics/TargetStatistics.html

- Columns in green indicate structures determined that have less than 30% sequence identity with any 20 residue (or longer) structure deposited in the Protein Data Bank (PDB).

- Nearly 3,000 structures fit this criterion have been determined by Structural Genomics

- PDB has in total around 40,000 entries.

- (Genbank has around 20,000,000 entries.)



Identification of DNA-binding proteins: 2 strategies

- Develop a training set (DNA-binding proteins and non-DNA-binding proteins

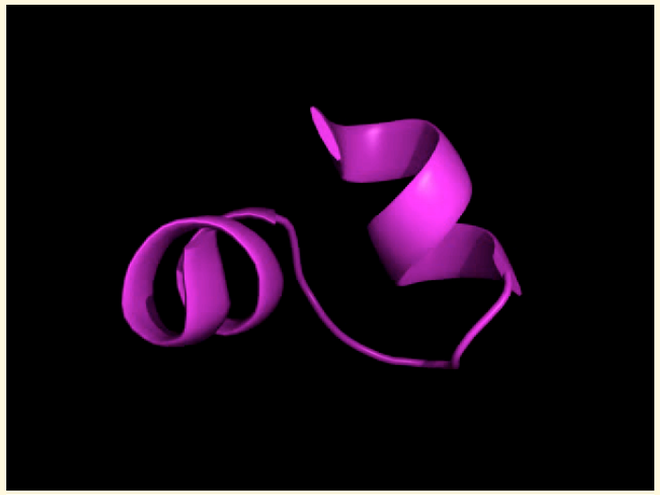
- Strategy 1: For each protein in training set gather a wide variety of local parameters associated with the proteins surface and use these to train a machine learning algorithm (Neural Network, SVM etc.)

    - Stawiski, Gregoret, Mandel-Gutfreund J. Mol. Biol. (2003) **326**, 1065-1079.

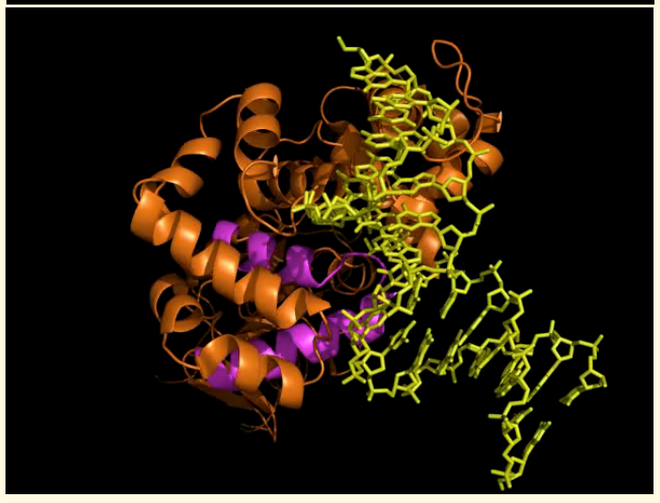
- Strategy 2: make use of the fact that as DNA-binding proteins often bind using a small set of structural motifs. Identify the motifs and use the RMSD of a given proteins against these motifs as an identification process.

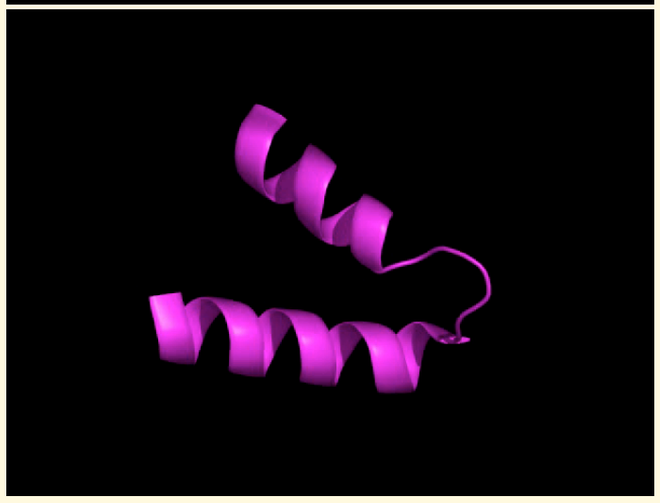
DNA-binding structural motifs

- Helix-Turn-Helix (HTH) motif

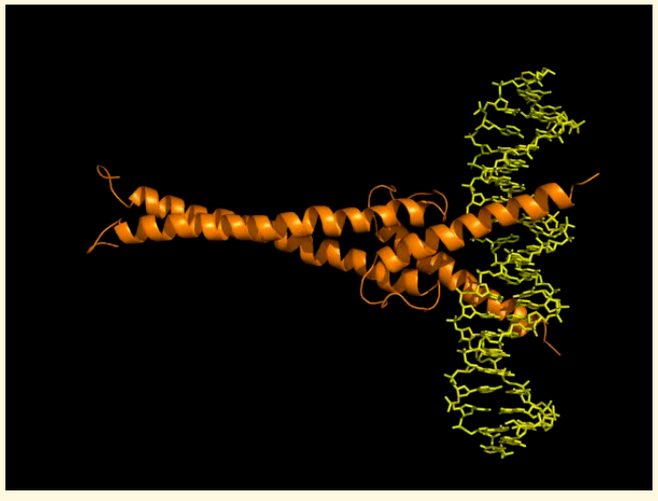


- Helix-hairpin-Helix (HhH) motif





- Helix-loop-Helix (HLH) motif



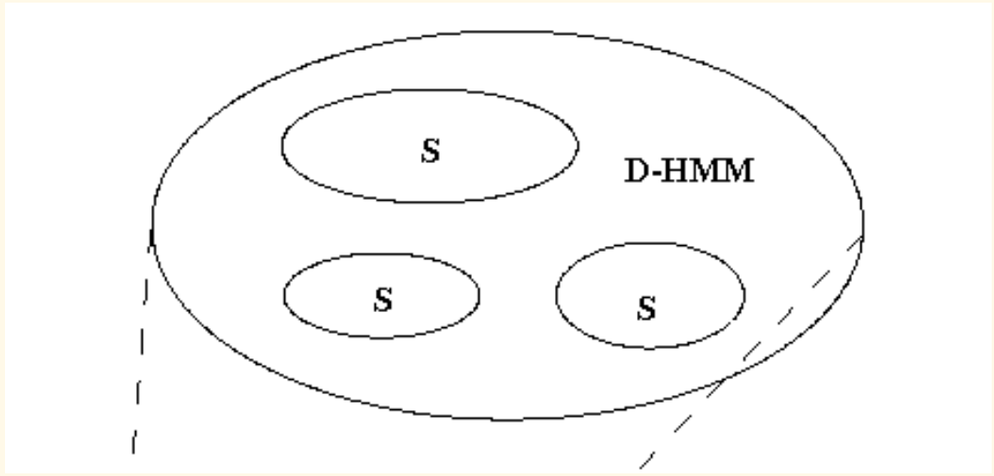
Overview of “families”

- S sequence family (35% sequence identity)

- Part of Domain HMM family

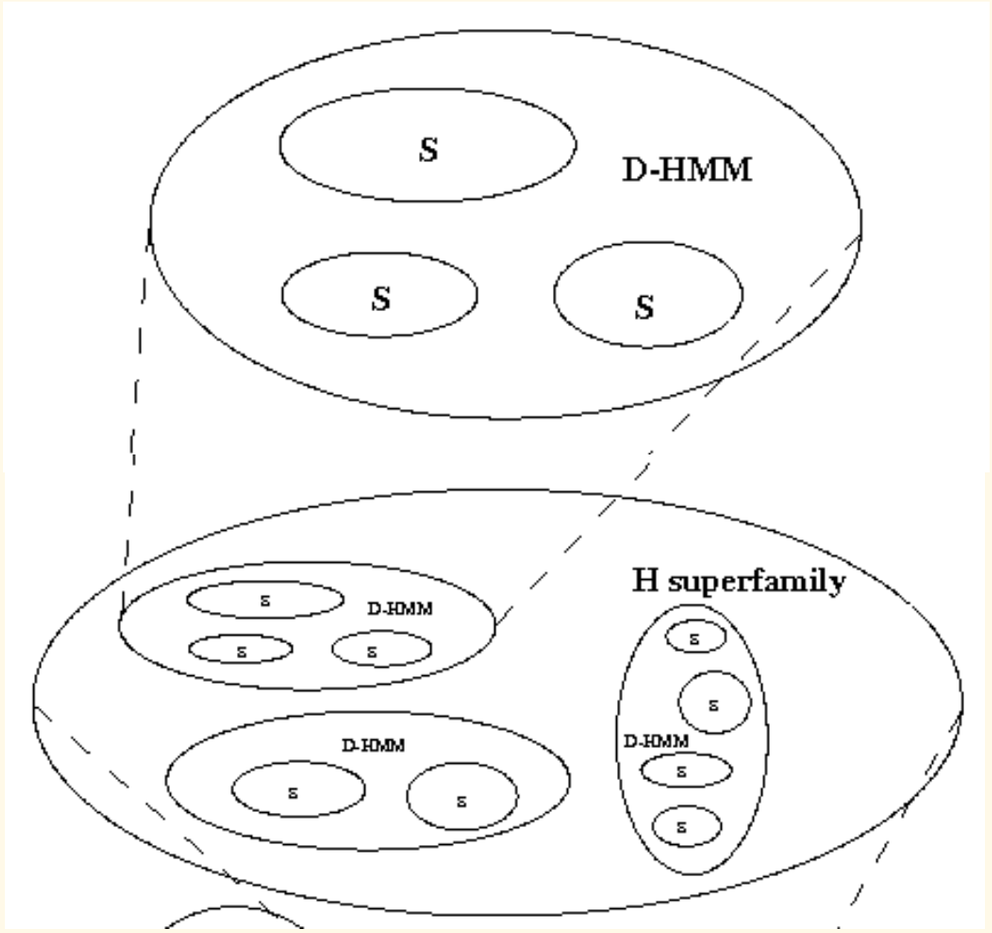
Overview of “families”

- S sequence family (35% sequence identity)
- Part of Domain HMM family



Overview of “families”

- S sequence family (35% sequence identity)
- Part of Domain HMM family
- Each D-HMM family is part of H superfamily (CATH, SCOP, etc.) (very distant homologues)



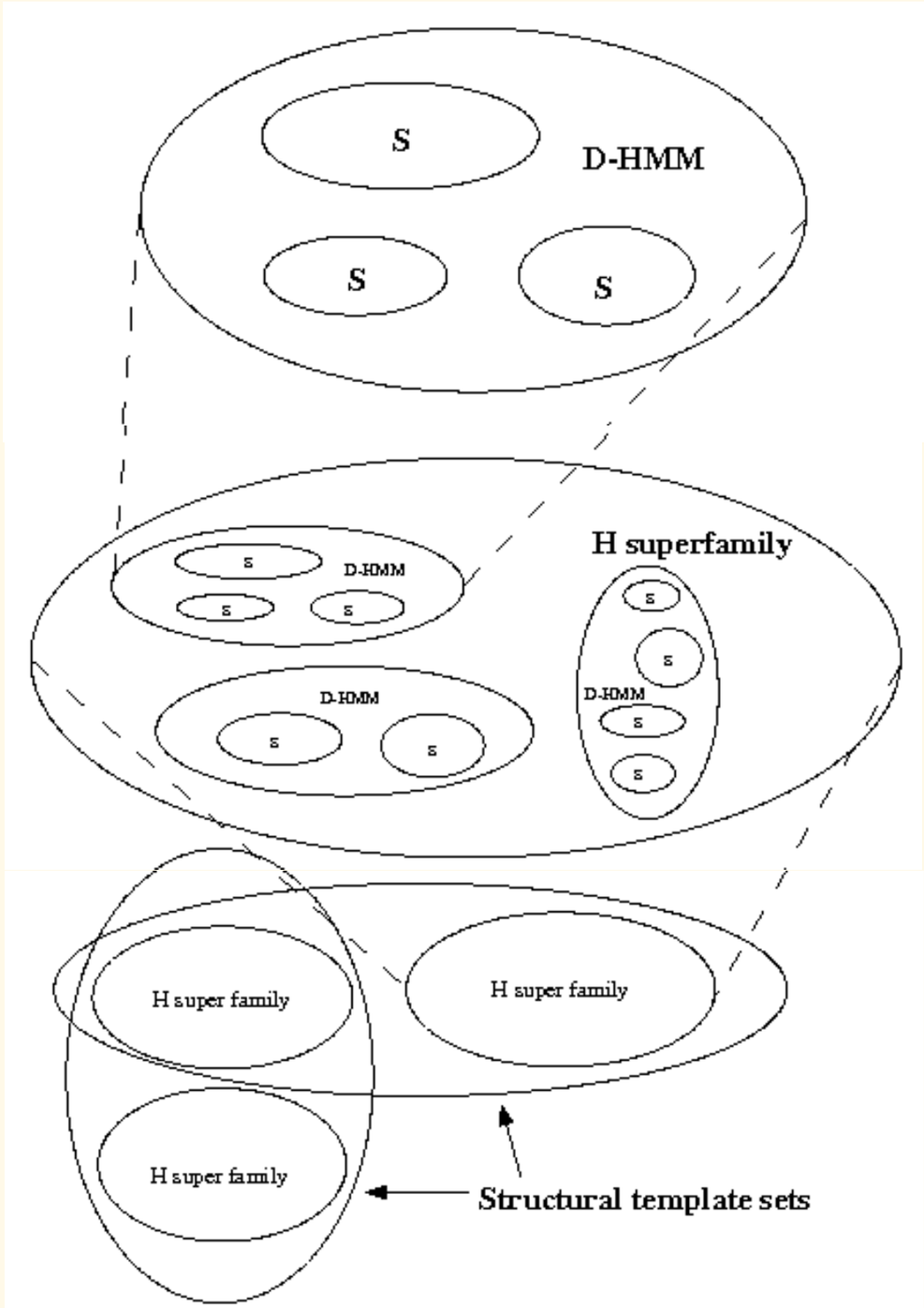
Overview of “families”

- S sequence family (35% sequence identity)

- Part of Domain HMM family

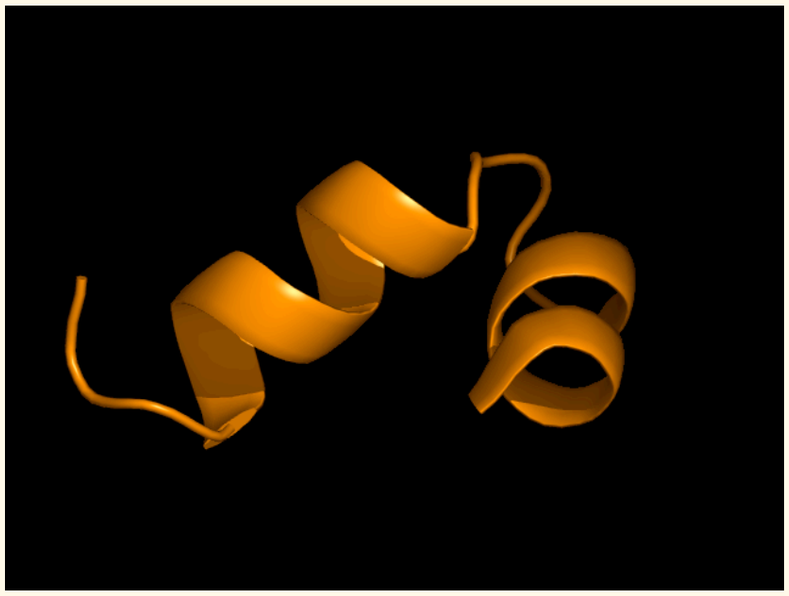
- Each D-HMM family is part of H superfamily (CATH, SCOP, etc.) (very distant homologues)

- H super-families can be match using structural templates (convergent evolution; plain old Physics/Chemistry)



Structural template

- Cα positions for a contiguous segment of a given protein spanning the structural motif.

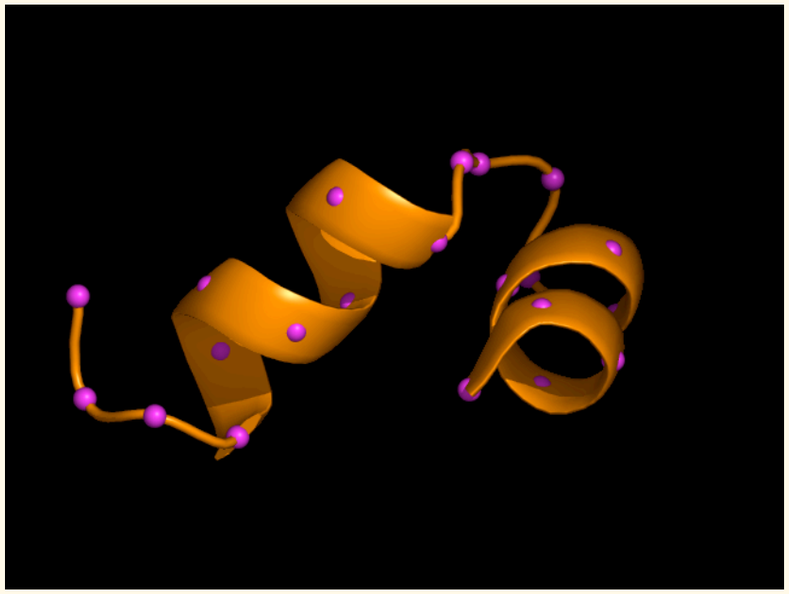


- Algorithm to identify minimum RMSD simply means sliding template along the Cα positions of given query structure - time is O(Nseq/Ntemplate).

- Can scan entire PDB in less than an hour.

Structural template

- Cα positions for a contiguous segment of a given protein spanning the structural motif.



- Algorithm to identify minimum RMSD simply means sliding template along the Cα positions of given query structure - time is O(Nseq/Ntemplate).

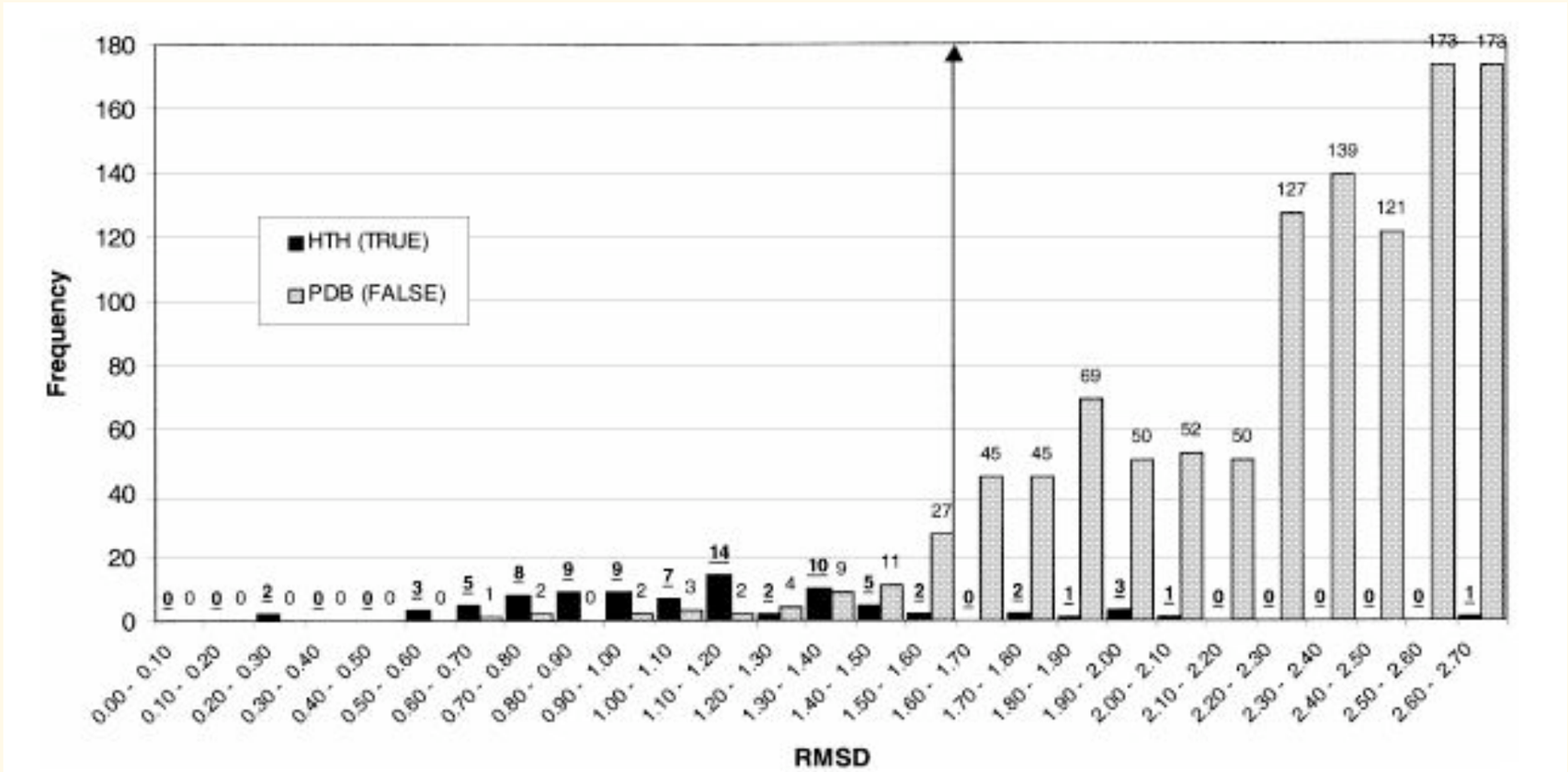
- Can scan entire PDB in less than an hour.

Helix-Turn-Helix structural motif

- Initial focus of study.
- Occur in 1/3 of known structural DNA-binding protein families.
- Examples of families with HTH motif:
    - Homeodomain (Drosophila development)
    - TFIIB (RNA polymerase promotors)
    - Interferon Regulatory Factors (IRF)
- C.F. http://www.biochem.ucl.ac.uk/bsm/prot_dna/prot_dna_cover.html

Results for HTH Structural motif

S. Jones *et al.* NAR, (2003), 31 (11), 2811-2823



- Initial set of 86 HTH proteins
- 76/86 True proteins identified
- 61/8264 of non-DNA binding set falsely identified.

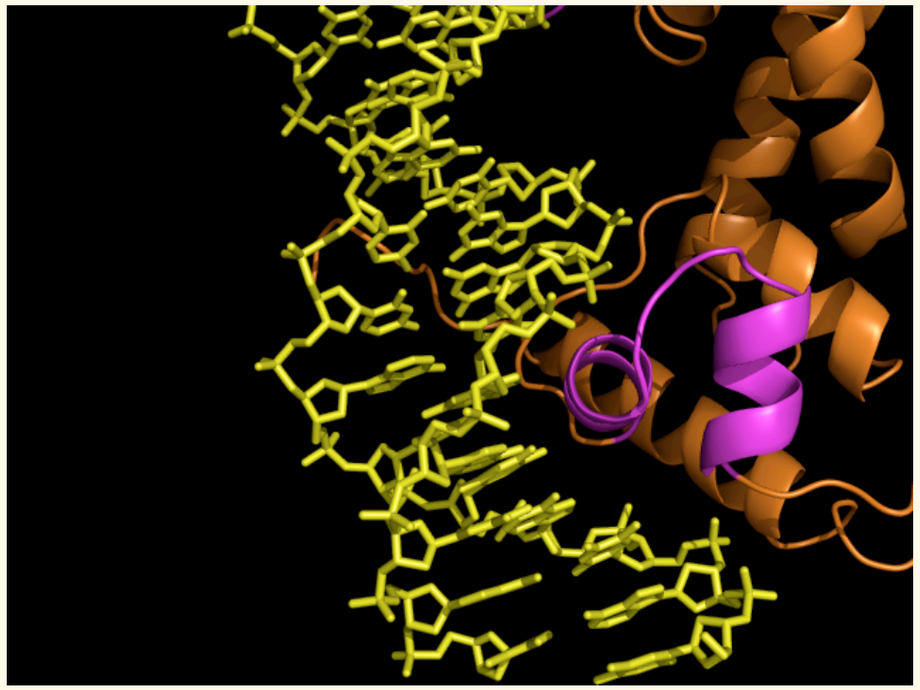
An extra condition: accessibility

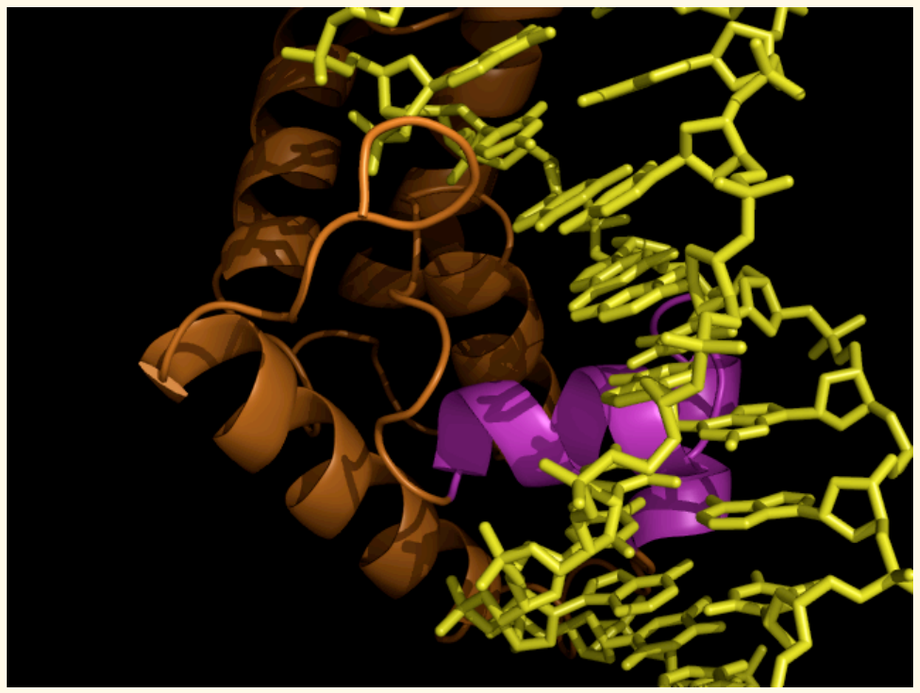
- A motif that binds should be solvent accessible.
- Introduced extra parameter, ASA.
- Set minimum ASA of putative motif to be 990 $Å^2$.
- Reduced number of false positives from 61 to 38.
- Nonetheless, it is worrying that a comparable absolute number of false positives exist.

The electrostatic potential

- Region around bases of DNA has a known negative charge.
- It seems reasonable to think that a protein that binds to such regions will have an overall positive charge.
- Example :- Tubby protein.
- Using charges distributed over a patch would appear to be the best method for elucidating this, but is more tricky than one might think.

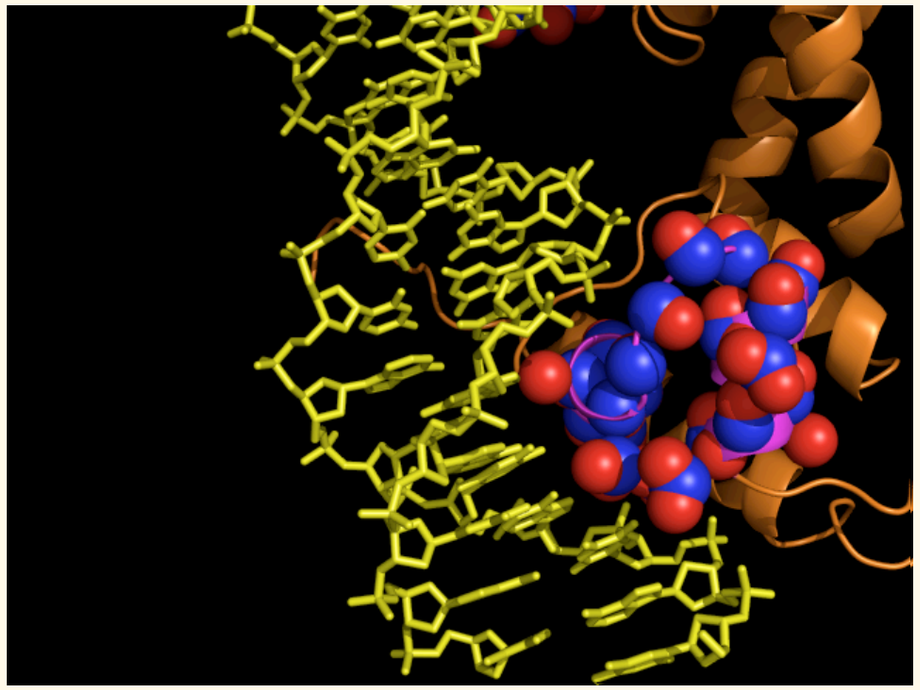
Electrostatic charge against potential

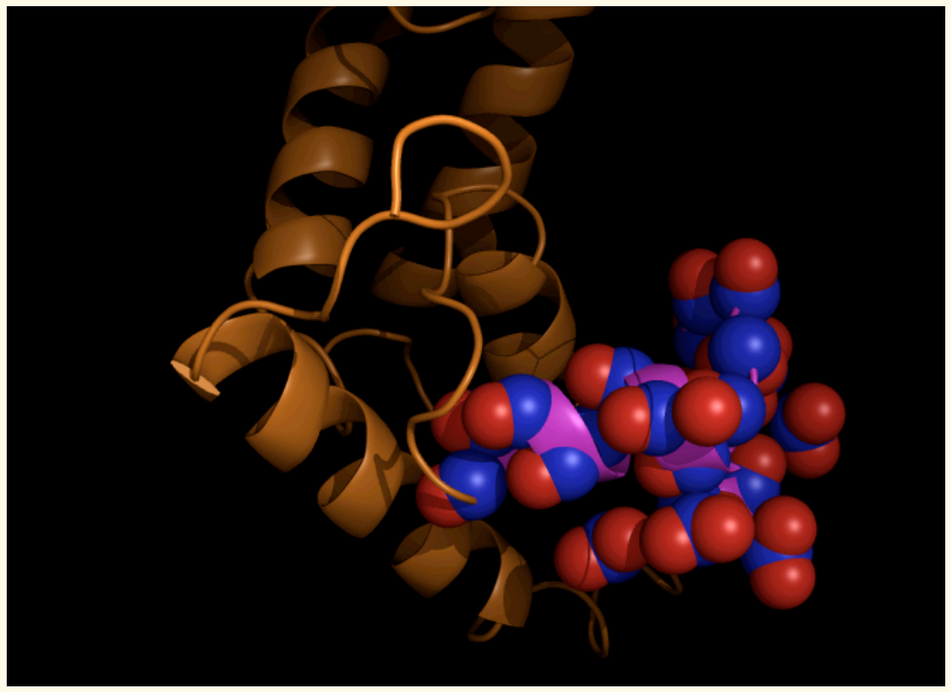




Electrostatic charge against potential

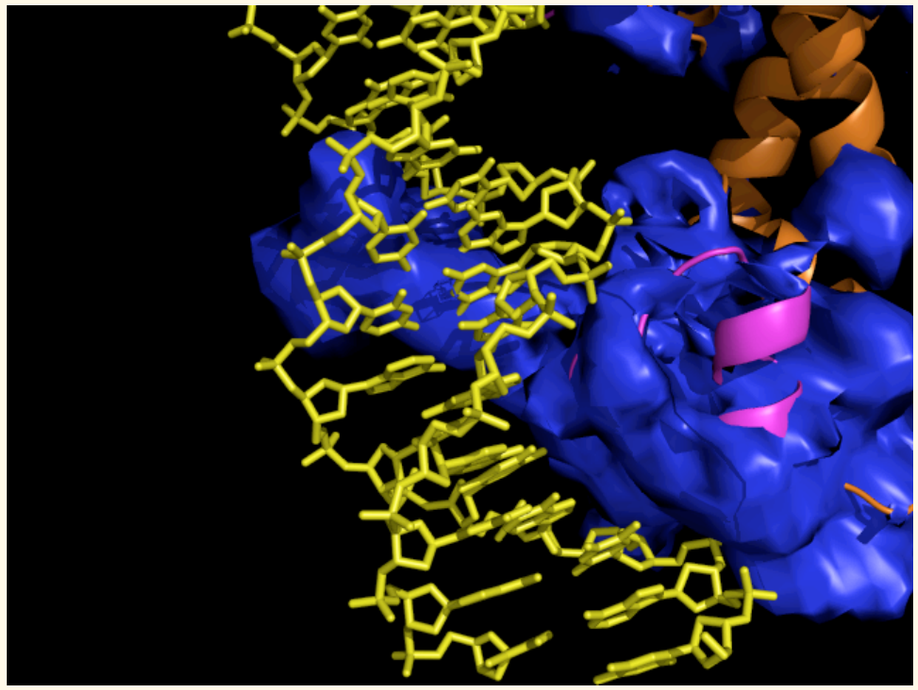
- Use partial charges initially



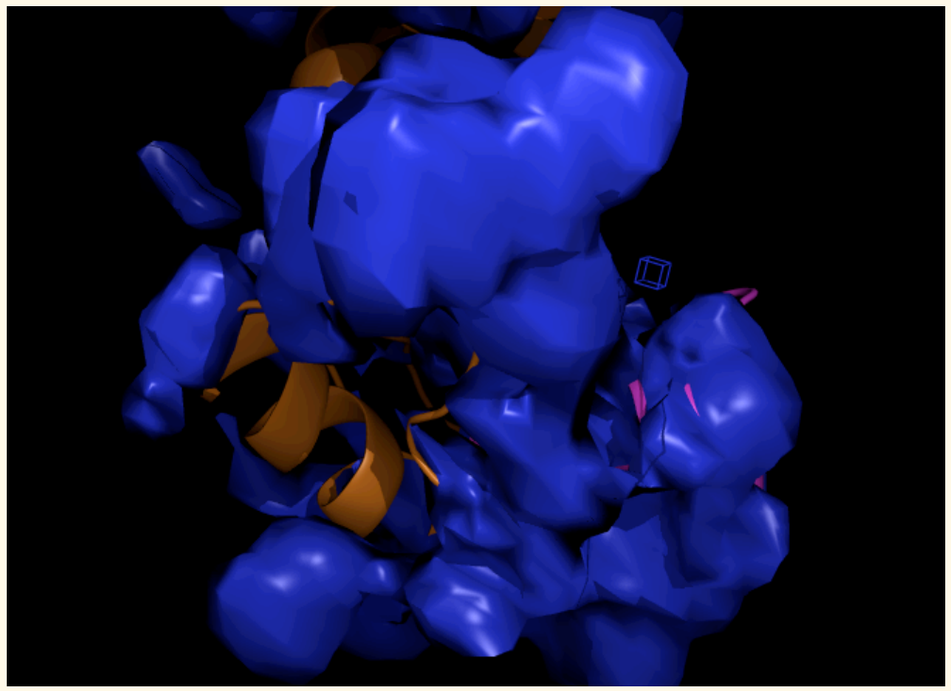


Electrostatic charge against potential

- Use partial charges initially



- isosurface +5 KeV electrostatic potential



Electrostatic Motif Score

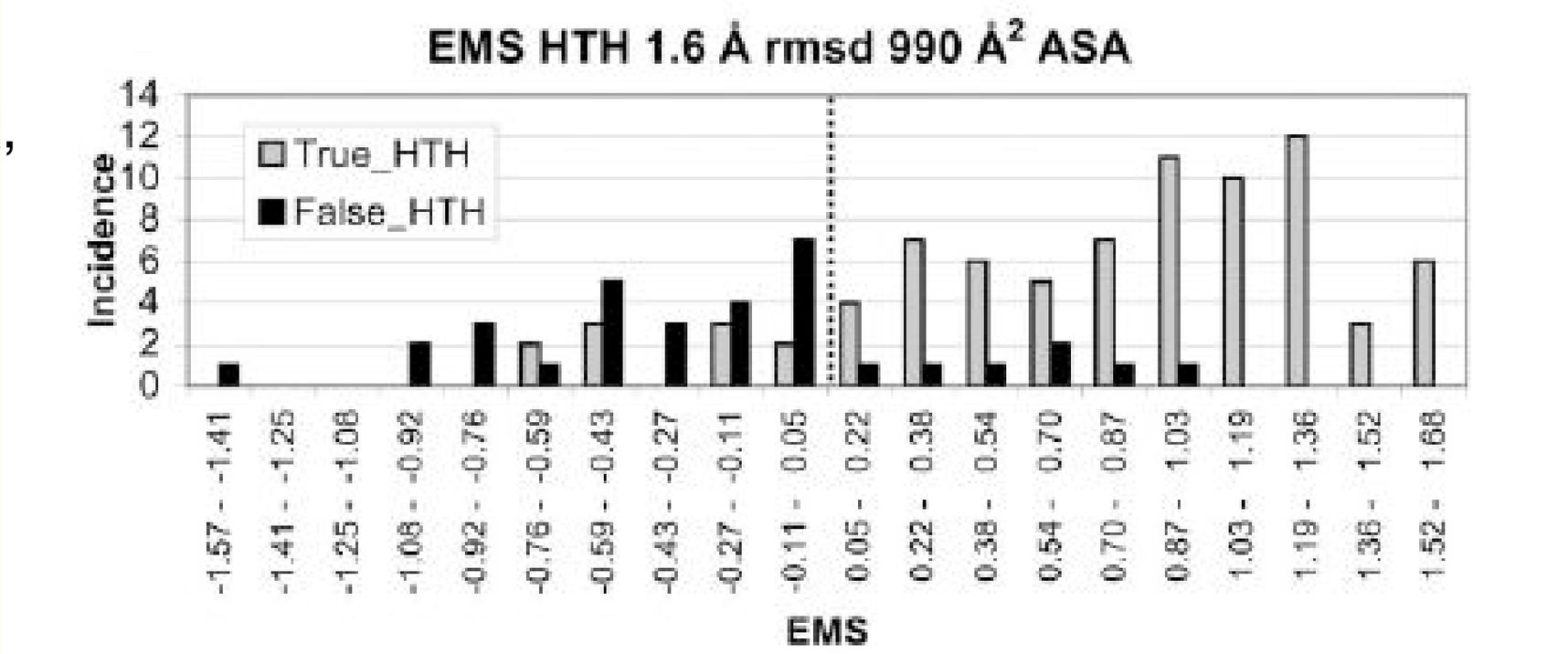
- Potential computed using Delphi (using a reduced charge set, i.e. not introducing Hydrogens)

- In order to associate potential with the surface of the motif, compute the following score :

$$
\begin{aligned}
EMS &= \frac{1}{N_M} \sum_{i \in M} \Delta Q_i \\
\Delta Q_i &= \frac{1}{\Delta S_i} \int_{\Delta S_i} V(\mathbf{r}) dA(\mathbf{r}) \\
N_m &= \text{Number of atoms on motif surface} \\
\Delta S_i &= 7\ \text{Å}\ \text{ exposed surface for ith atom}
\end{aligned}
$$

Addition of EMS

Shanahan *et al*, NAR, (2004), 32 (16), 4732-4741



- Number of false positives fell from 33 to 8

Putting it all together... HTHquery

- Analysis was so far very naive.
- Used Neural Network (Linear Predictor) for all three variables.
- Training set
  - 79 DNA-binding chains
  - 490 non-DNA-binding chains (RMSD < 2.5 Å)
- 7 structural templates

Ferrer-Costa et al. Bioinformatics, 21 (18), 2005, 3679-3680

How does it compare

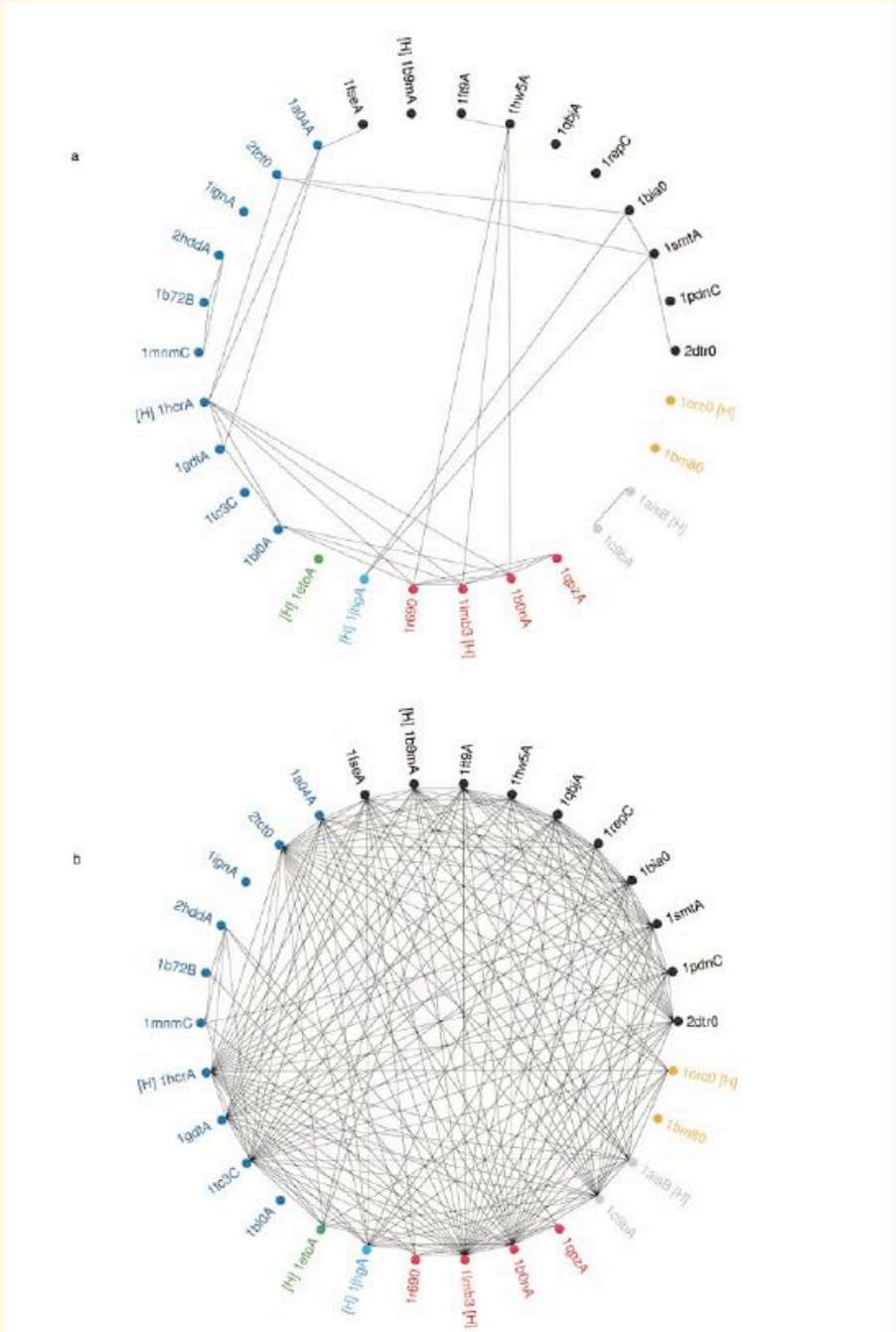
	HTHquery	Stawiski et al.
Sensitivity	83.5%	81%
Specificity	99.2%	94%

Motifs and convergent evolution

- Hierarchy of evolved structures.

- HMM’s can only usually pick out those representatives of its own sequence family.

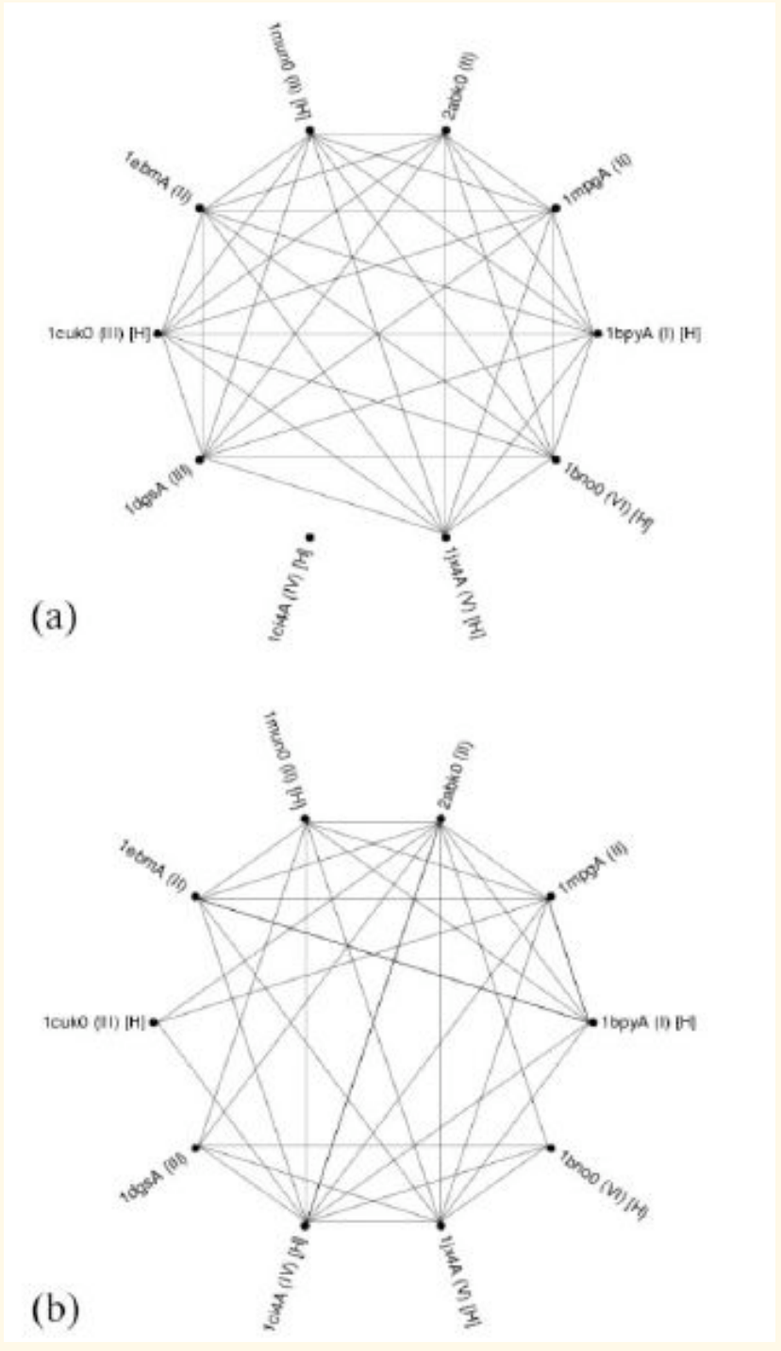
- The template approach can pick out representatives across seperately evolved structures.



The other motifs..

- HhH motif can be picked out almost as well using HMM.

- HLH proteins all sit in the same sequence family - remarkable given how important the HLH transcription factor is in Drosophila and Arabidopsis.



Thinking aloud

- The methods discussed here appear to work well in the case of the Helix-Turn-Helix motif, where plenty of examples exist of convergent evolution.

- Comparatively it is much easier to just use HMM's to identify proteins with a HhH or HLH.

- The idea of structural templates are used extensively in inferring enzymatic function and indirectly protein-protein interactions.

- In the former case in particular, if one could determine the level of convergent evolution for a given enzymatic cleft one could then focus on those that exhibit such behaviour and leave the others to sequence based approaches.

Further issues

- Where are the Zinc-Fingers ?
    - Too structurally variable !
    - At present Stawiski approach is probably the best...

- HTH method still needs improvement. IN particular many false negatives occur because wrong motif is identified.
- Many DNA-binding proteins are disordered until they bind to DNA.
- Will this obviate this kind of work ?

Acknowledgments

- Jonathan Barker, European Bioinformatics Institute

- Carles Ferrer-Costa, Bioinformatics, UB, Barcelona

- Mario Garcia, Departmento de Ciencias Quimicas, Universidad San Pablo CEU, Madrid.

- Sue Jones, Dept. of Biochemistry, University of Sussex

- Janet Thornton, European Bioinformatics Institute

Thank you for your time !