Development of technologies for large-scale knowledge-engineering in the post-genomic era

Stephen Muggleton
Department of Computing
Imperial College, London

October, 2003

Overview

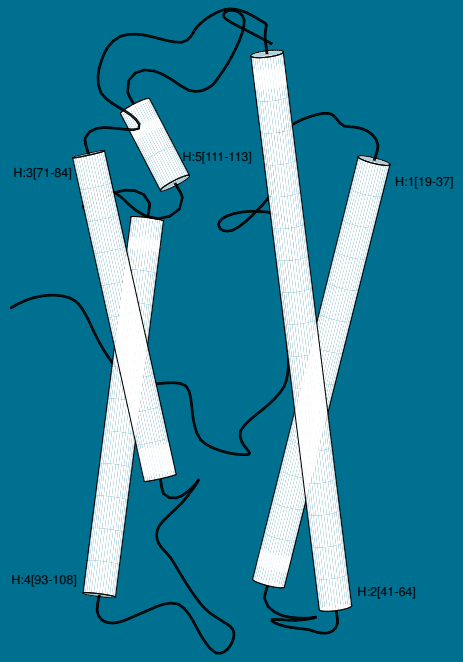
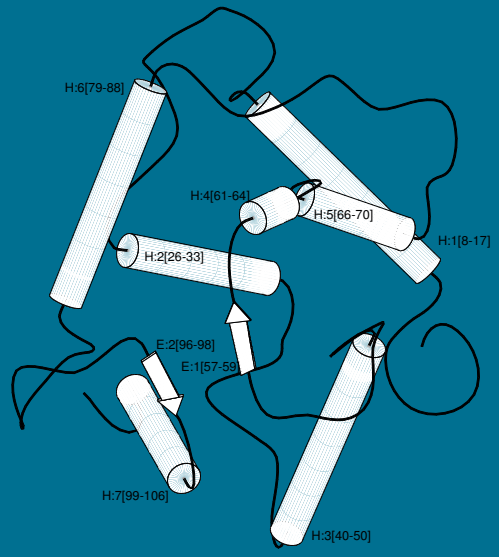
Inductive Logic Programming

Previous Biological and Chemical domain results

Ongoing research

Conclusions

Proteins in fold class
“4-helical up-and-down bundle”

Positive(12)	Negative(12)
	

2mhr - Four-helical up-and-down bundle

1omd - EF-Hand

Declarative representation: logic programs

```
fold('Four-helical up-and-down bundle',P) :-
    helix(P,H1),
    length(H1,hi),
    position(P,H1,Pos),
    interval(1≤ Pos ≤ 3),
    adjacent(P,H1,H2),
    helix(P,H2).
```

The protein P has fold class "Four-helical up-and-down bundle" if it contains a long helix H1 at a secondary structure position between 1 and 3, and H1 is followed by a second helix H2.

Inductive logic programming

Background knowledge:	B
Examples:	$E = E^+ \wedge E^-$
Hypothesis:	H
Prior consistency:	$B \wedge E^- \not\models \square$
Prior necessity:	$B \not\models E^+$
Posterior consistency:	$B \wedge H \wedge E^- \not\models \square$
Posterior sufficiency:	$B \wedge H \models E^+$
	$B \wedge \overline{E}^+ \models \overline{H}$

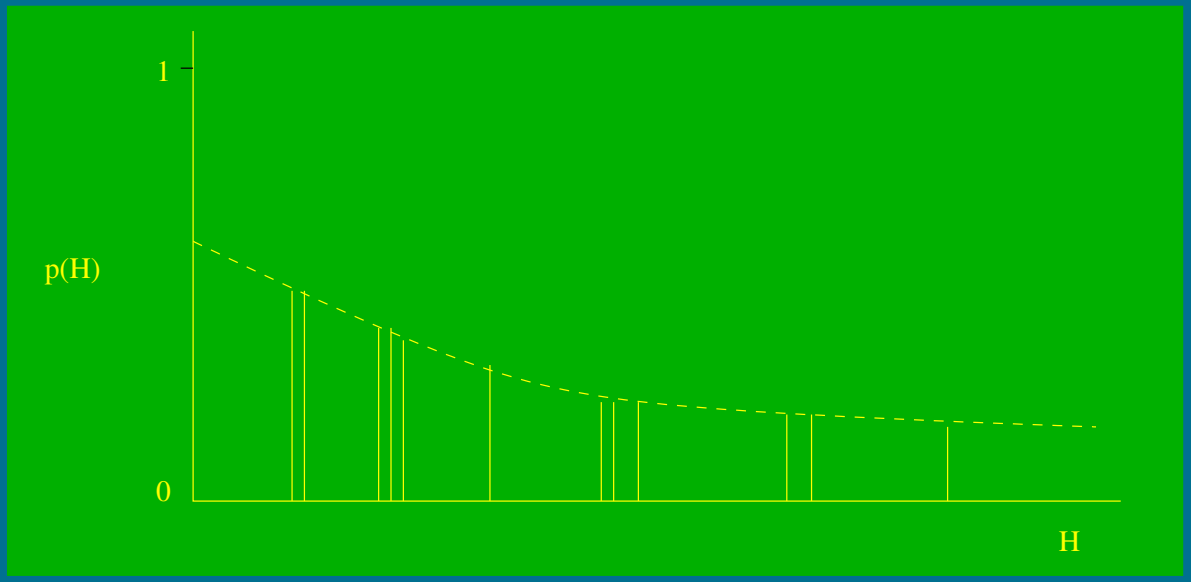
Inductive logic programming

Background knowledge. Protein sequence, partial grammar, domain constraints.

Examples. Molecules, annotated sentences.

Hypothesis. Explanation of molecular 3-D shape, new clauses in a grammar.

Bayes' framework



Maximise $p(H|E)$ for which

$$H \models E \text{ and } H \text{ consistent.}$$

Applied to analysis of positive-only learning, relevance, predicate invention and learning stochastic logic programs.

Discovery of biological function

Biological functions regulated by docking of small molecules (ligands) with sites on large molecules (proteins).

Drugs (eg. beta-blockers) mimic natural small molecules (eg. adrenaline).

Effectiveness of drugs depends on correct shape and charge distribution.

IC/York/Oxford/ICRF/SB ILP biological results

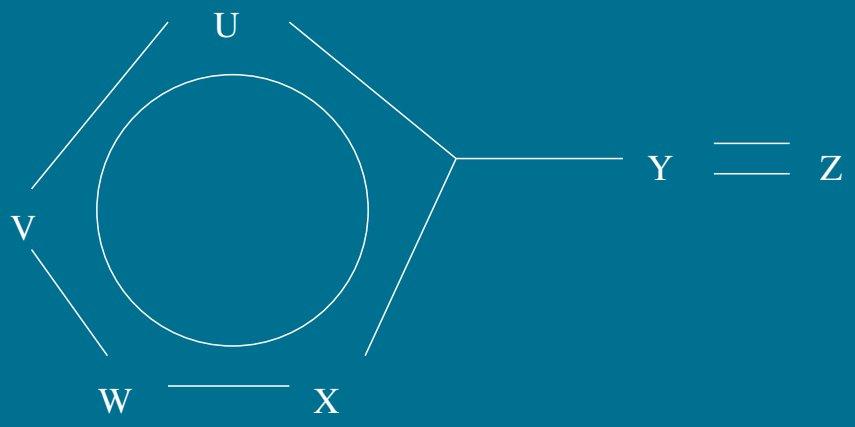
- Protein secondary-structure (Pr. Eng. 1992).
- "1D" structure-activity (PNAS 1992).
- Beta-strand arrangement (PTRS 1994).
- "2D" prediction of mutagenesis (PNAS 1996).
- "3D" prediction of pharmacophore (MLJ, 1998).
- 23 protein folds, ave. accuracy 82% (50% default) (MLJ, 2001; JMB, 2001).
- Neuropeptides, 108 times SB recognition rate (JCB, 2001).
- Active ML, metabolic pathways, (ETAI, 2001).
- SLP learning, metabolic pathways, (ETAI, 2002; ILP, 2003).

Mutagenesis

Related to carcinogenicity.

Measured by Ames test.

Not always determinable empirically.



New structural alert (42 molecules):

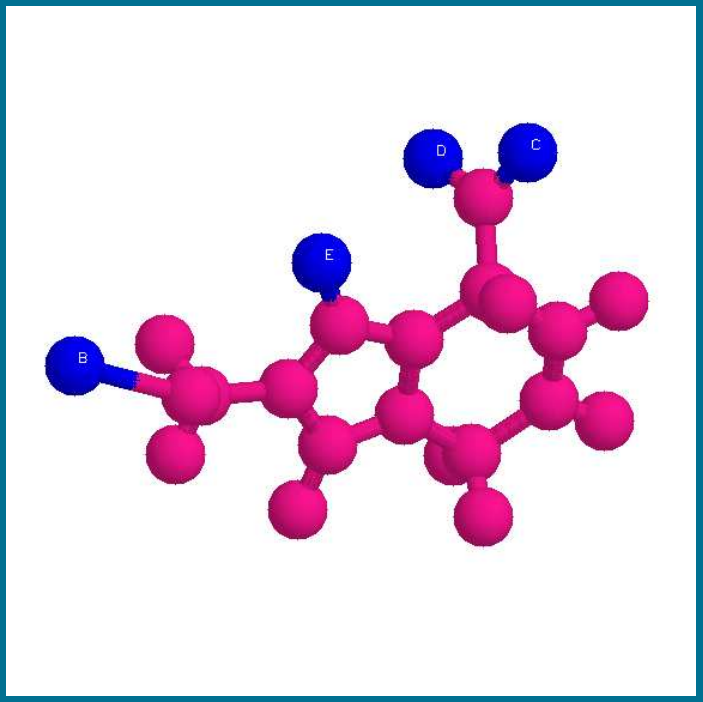
Mutagenic if there exists a double bond

connected to a 5-aromatic ring.

PTE carcinogenesis competition

Method	Accuracy	P
Ashby	0.77	0.29
Progol	0.72	1.00
RASH	0.72	0.39
TIPT	0.67	0.11
Bakale	0.63	0.09
Benigni	0.62	0.02
DEREK	0.57	0.02
TOPKAT	0.54	0.03
CASE	0.54	< 0.01
COMPACT	0.54	0.01
Default	0.51	0.01

Pharmacophores



Geometric arrangement of key features involved in ligand-protein binding.

Hypotheses

Molecule A is an ACE inhibitor if:

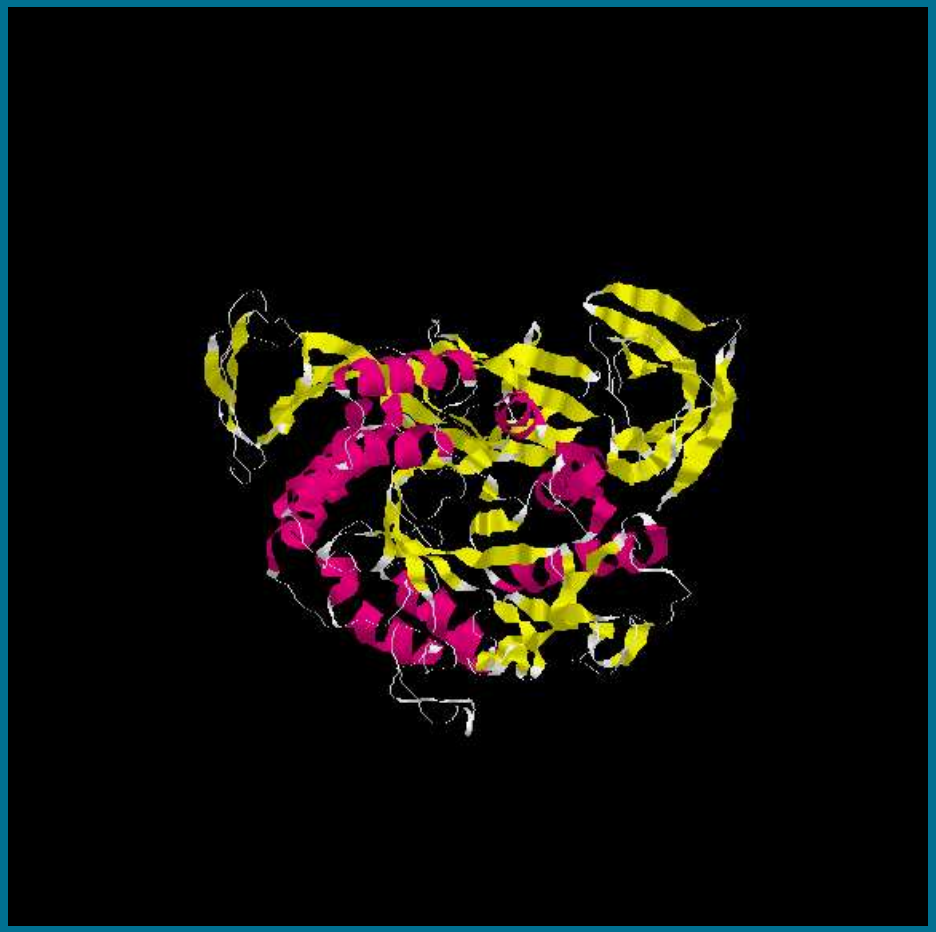
molecule A can bind to zinc at a site B, and

molecule A contains a hydrogen acceptor C, and

the distance between B and C is 7.9 +/- 1.0 Angstroms, and

...

TIM-barrel fold



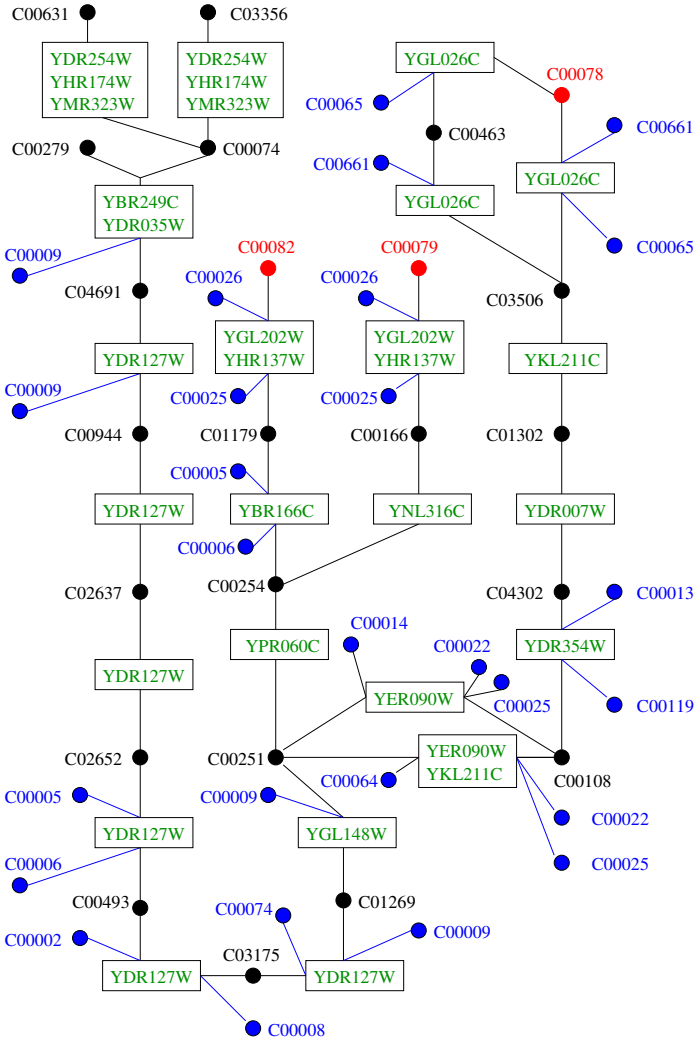
Fold-prediction results

Class	**Fold**	**Feature-based** (%)	**Relational** (%)
All-α	3-helical	82.3	84.6
	EF-hand	71.4	82.2
All-β	OB-fold	75.0	80.39
α/β	P-loop	69.2	79.2
	α/β hydrolases	85.7	85.9
$\alpha+\beta$	β-Grasp	66.3	79.1
	Ferre.-like	73.6	81.1
Overall		75.1 $\pm$ 1.6	82.1 $\pm$ 1.4

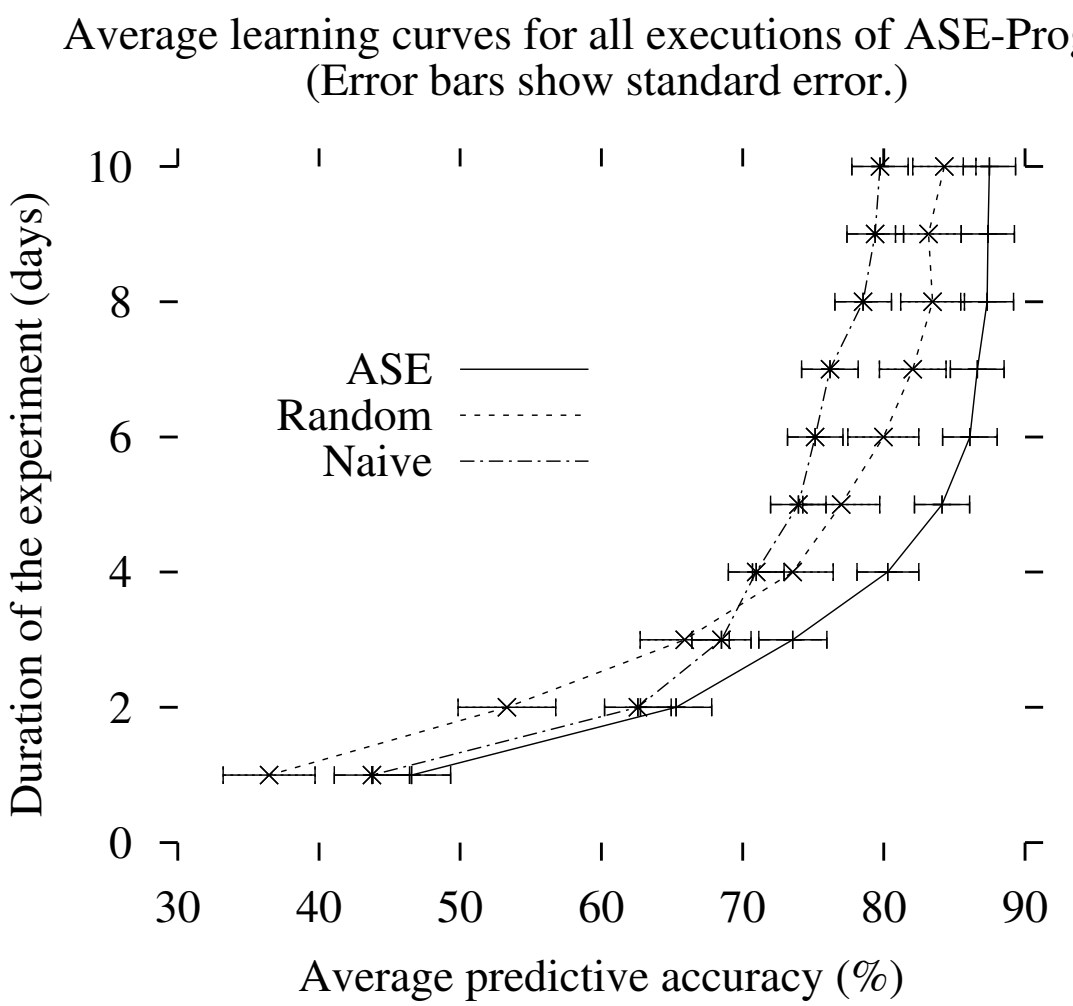
Neuropeptide multi-strategy learning

Amalgam	*RA*
Only proportions	0
Only Length	1.6
Only SignalP	11.7
Only Grammar	10.8
Length + SignalP + Grammar	0
Proportions + Length + SignalP	29.2
Proportions + Length + Grammar	33.2
Proportions + SignalP + Grammar	15.0
Proportions + Length + SignalP + Grammar	107.7

Aromatic amino acid pathways for yeast



Learning curves - time



Present research goals

Revision of large-scale hierarchical distributed knowledge-bases

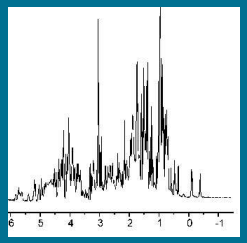
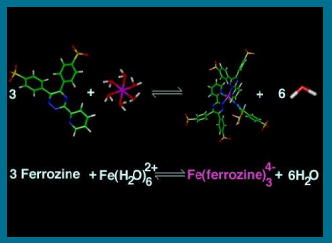
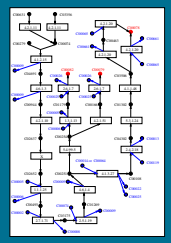
Dynamic structured models of molecules, cells and organisms

Revision using Machine Learning (ML) and Statistics

Incompleteness and imprecision require Uncertainty Reasoning (UR)

Combining ML/Stats models requires expressive framework for ML+UR.

Metalog project

	Time	Space	Action	Uncertainty
Logics	Temporal	Spatial	Action-based	Probabilistic
Application	Toxins 	Protein Structure 	Cell Biochemistry 	Pathways 

Conclusion

- LPs provide easily comprehended representation for discovery.
- ILP allows use of expert domain knowledge.
- Supports re-use of learned knowledge.
- Closed loop machine learning promises partial automisation of science.
- ILP allows humans to maintain window in the loop.
- PILP supports uncertainty.

Bibliography

ILP general

Muggleton (1999), *Inductive Logic Programming: issues, results and the LLL challenge*, Artificial Intelligence, 114(1):283-296.

Muggleton and De Raedt (1994), *Inductive Logic Programming: Theory and Methods*, Journal of Logic Programming, 19,20, 629-679.

ILP and Scientific Discovery

Muggleton (1999), *Scientific Knowledge Discovery using Inductive Logic Programming*, Communications of the ACM, 42(11):42-46.