Knowledge Discovery in Biological and Chemical domains

Stephen Muggleton
Department of Computer Science
University of York

February, 2001

Overview

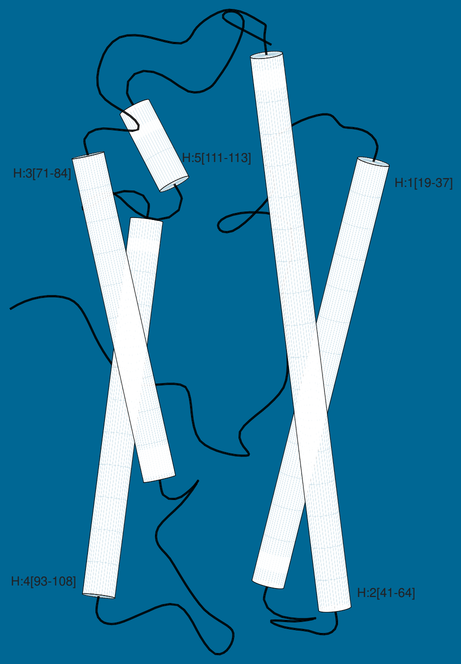
Inductive Logic Programming

Biological and Chemical domain results

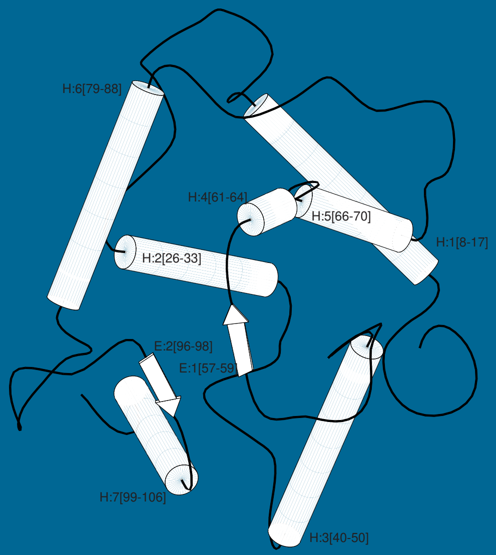
Conclusions

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.

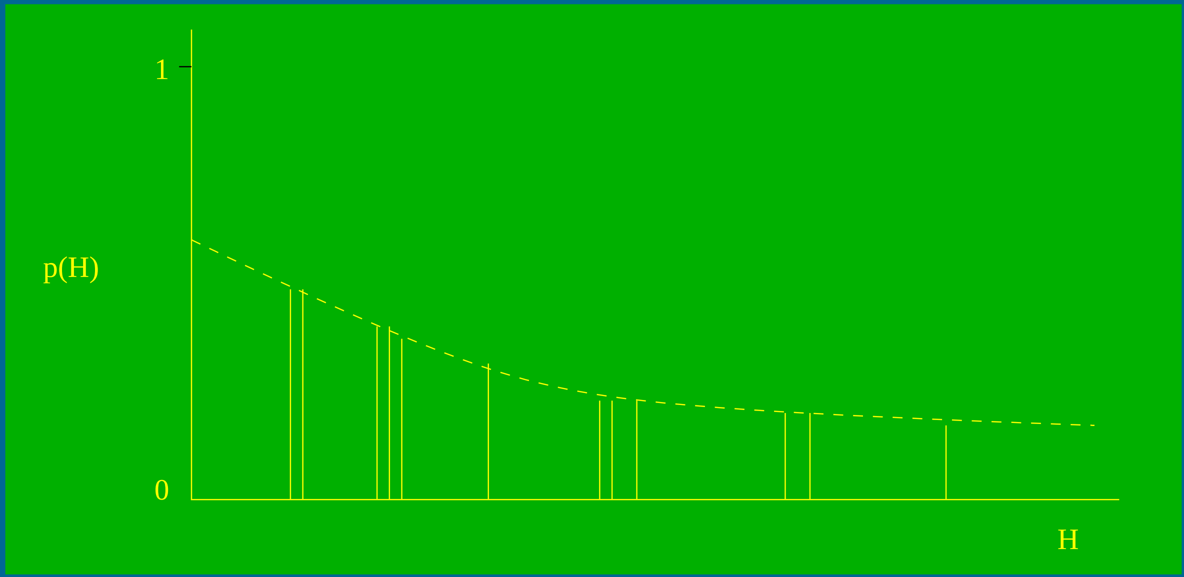
Inductive logic programming

Background knowledge. 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).

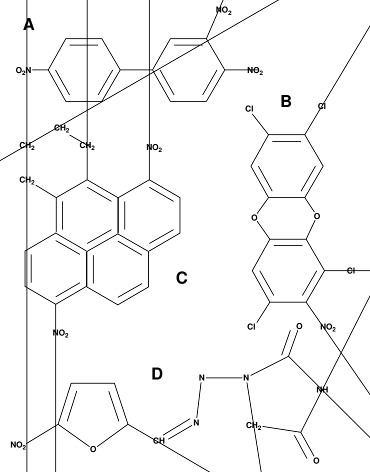
Drugs (eg. beta-blockers) mimic natural small molecules (eg. adrenaline).

Effectiveness of drugs depends on correct shape and charge distribution.

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 (ICML, 2000).

Mutagenesis examples



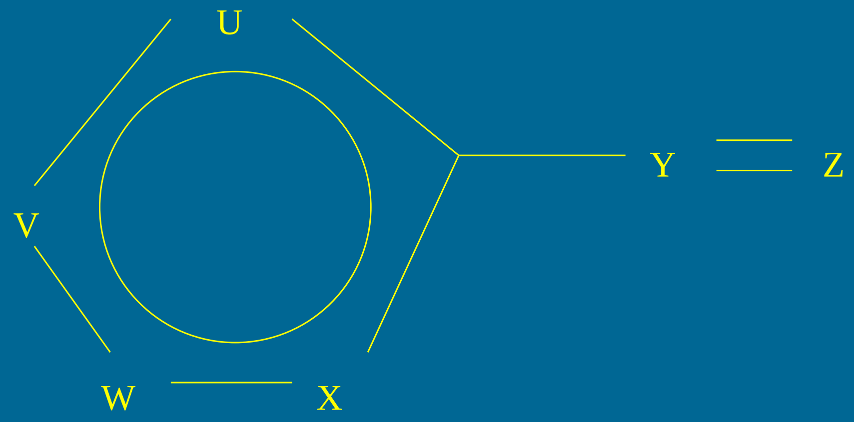
Atoms	Bond
atom(m10,a5,o).	bond(m10,a5,a6,2).
atom(m10,a6,c).	bond(m10,a6,a9,1).

Mutagenesis result

Related to carcinogenicity.

Measured by Ames test.

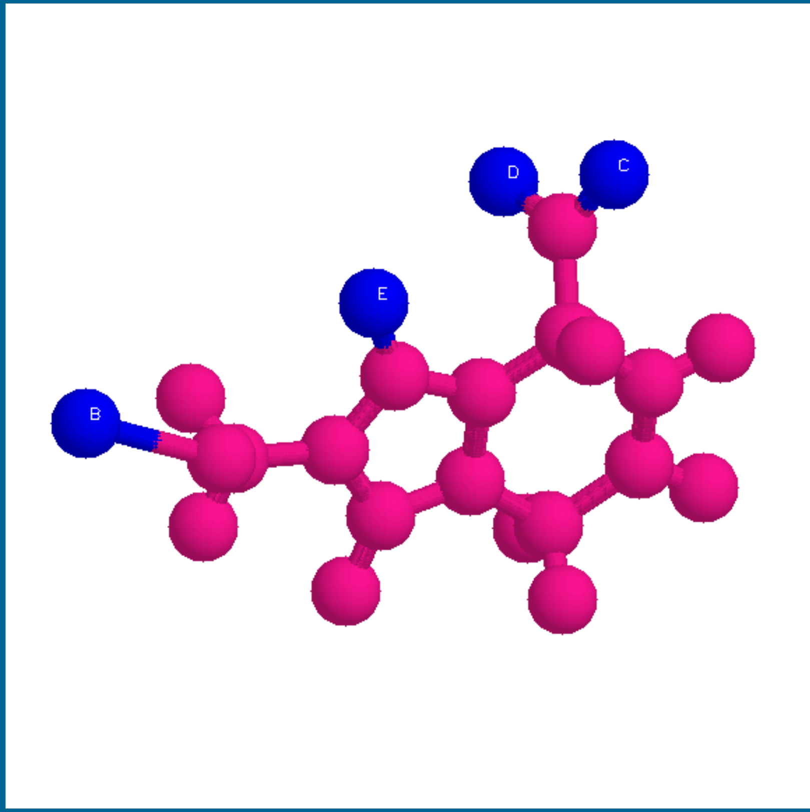
Not always determinable empirically.



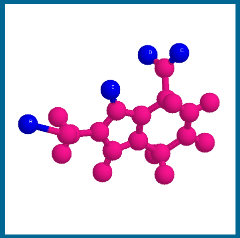
New structural alert (42 molecules):

Mutagenic if there exists a double bond connected to a 5-aromatic ring.

Pharmacophores



3D descriptions



Condensed	Uncondensed
atom(m1,a1,o,2,-3.43,-3.12,0.05).	in(m1,a1). elem(a1,o). sp_hybridized(a1,2). pos_x(a1,-3.43). pos_y(a1,-3.12). pos_z(a1,0.05).

Pharmacophore 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

. . .

Protein representations

Lists. alpha(p155C,1,[n,e,g,d,a,a,k,g,e,k,e,f,n,k,a,c,h,m,i]).

Indexed. position(p155C,1,n).

position(p155C,2,e).

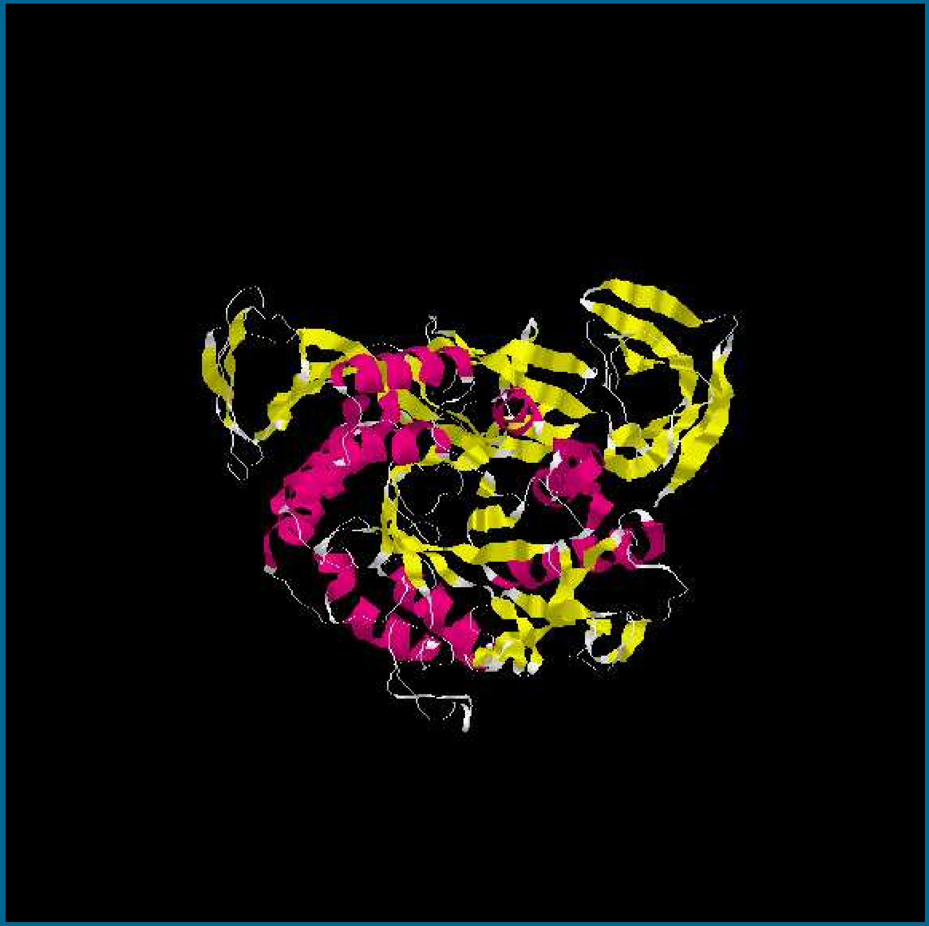
alpha(p155C,1).

alpha(p155C,2).

hydrophobic(h).

hydrophobic(w).

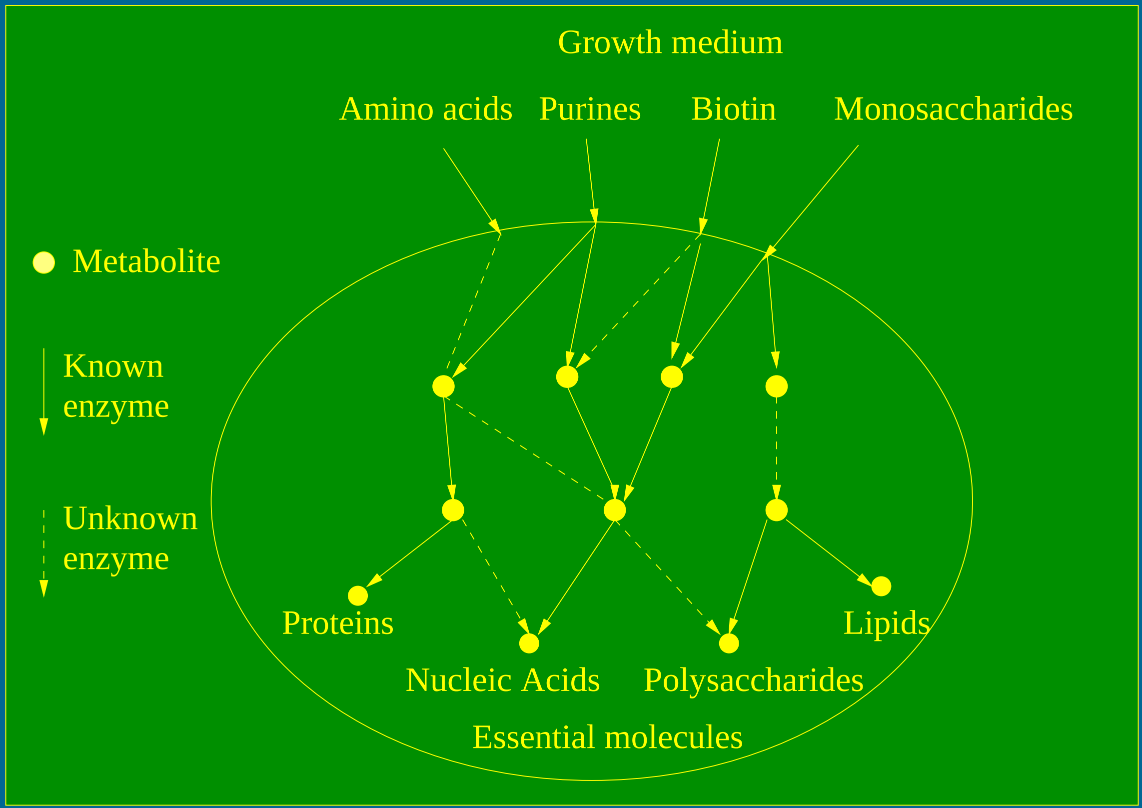
TIM-barrel fold



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

Functional genomics



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.

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.