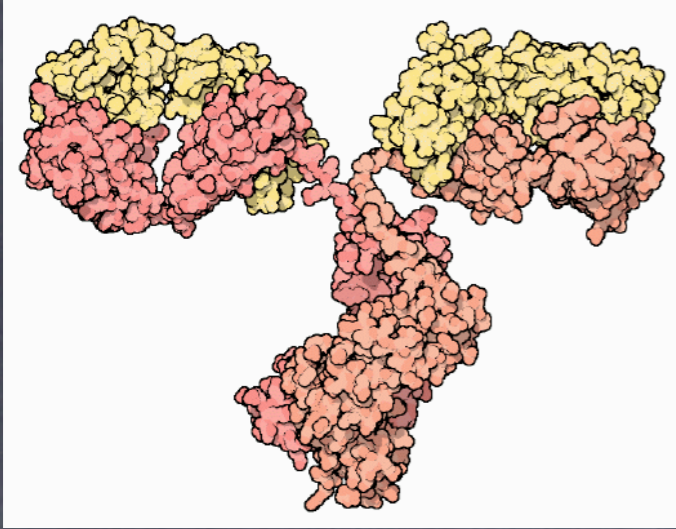


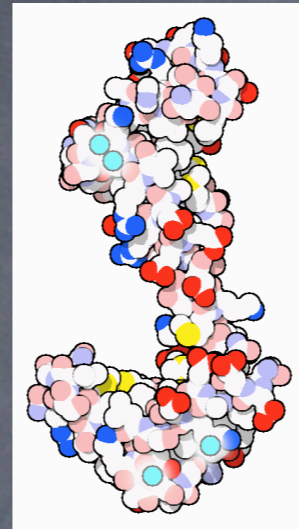
Automatic Derivation of Maximally Representative Database Subsets Based on Fragments

Jens Kleinjung
Division of Mathematical Biology
National Institute for Medical Research
London, UK

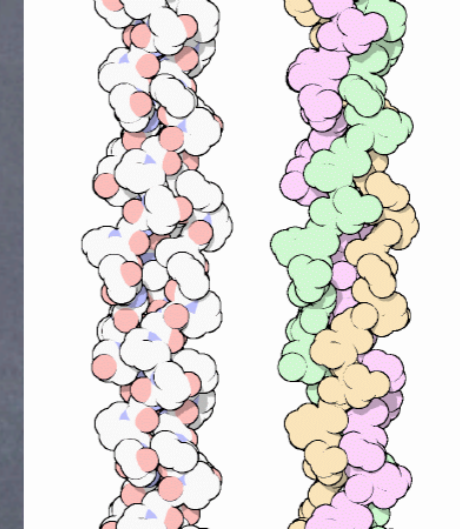
Proteins



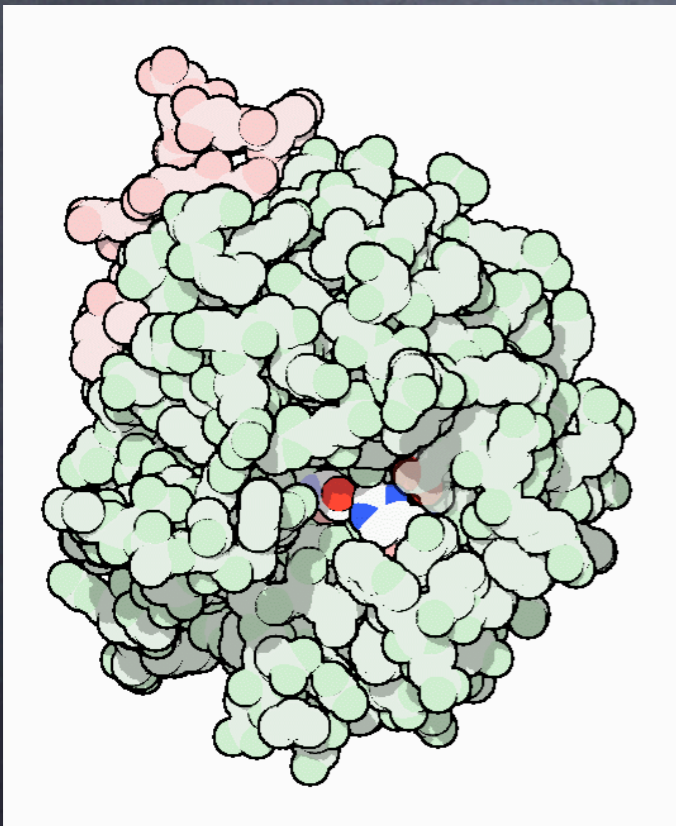
antibody



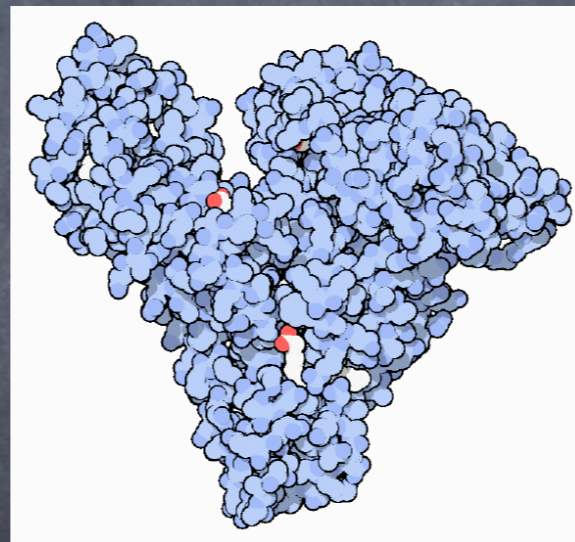
calmodulin



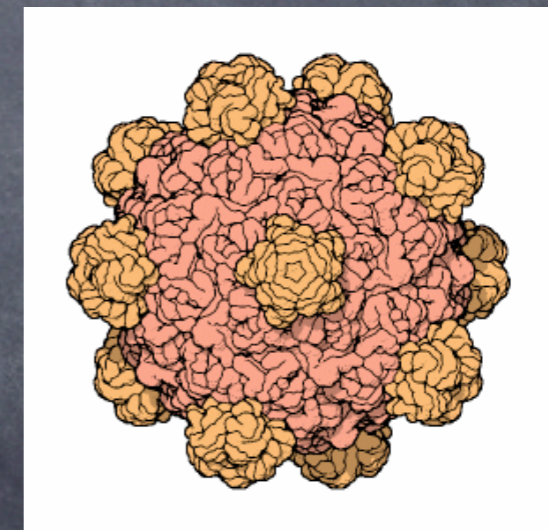
collagen



thrombin



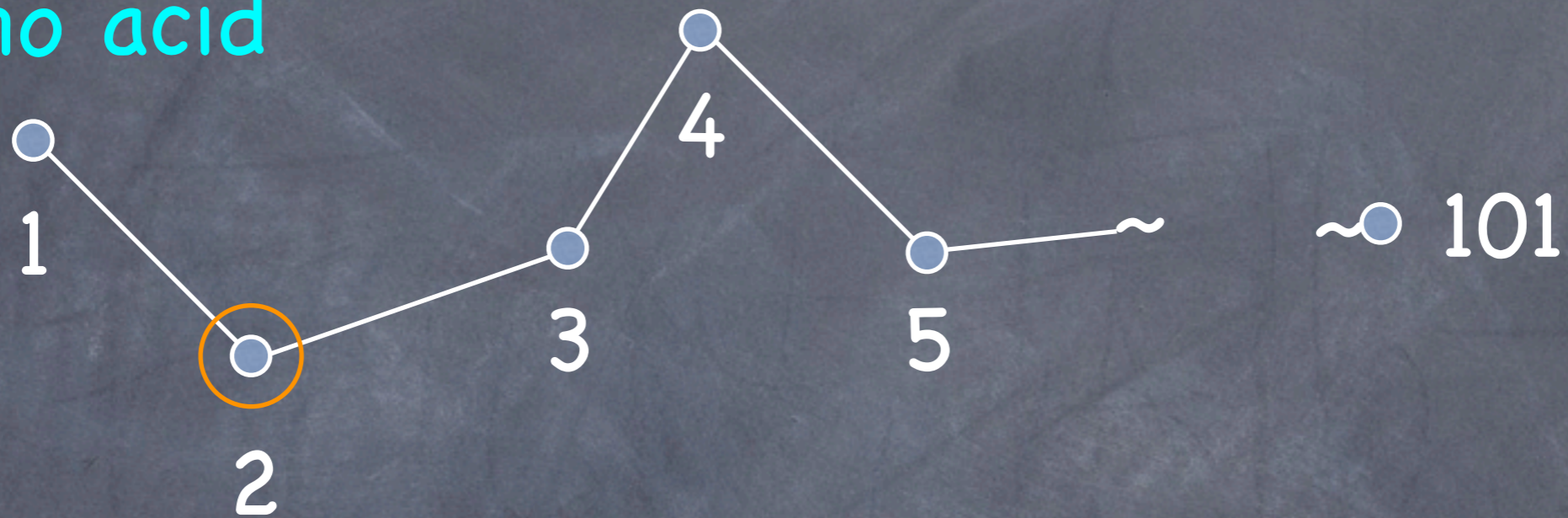
serum
albumin



phage

Conformational Space

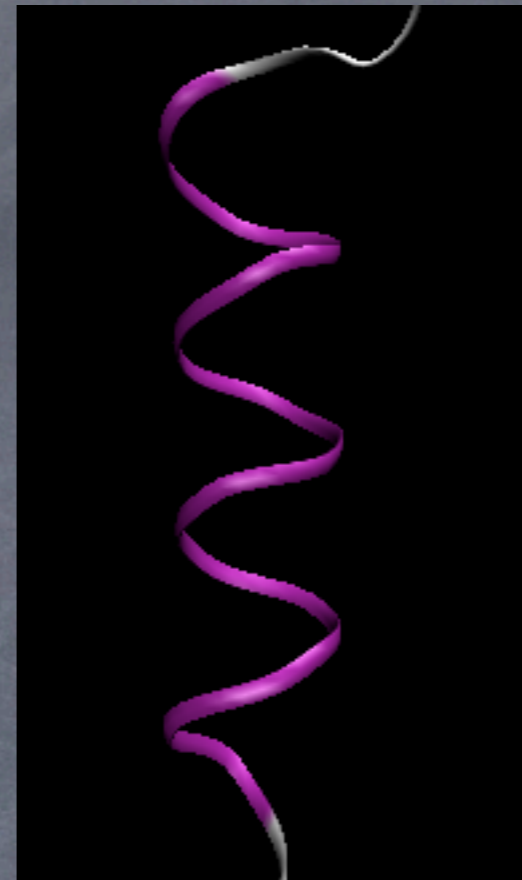
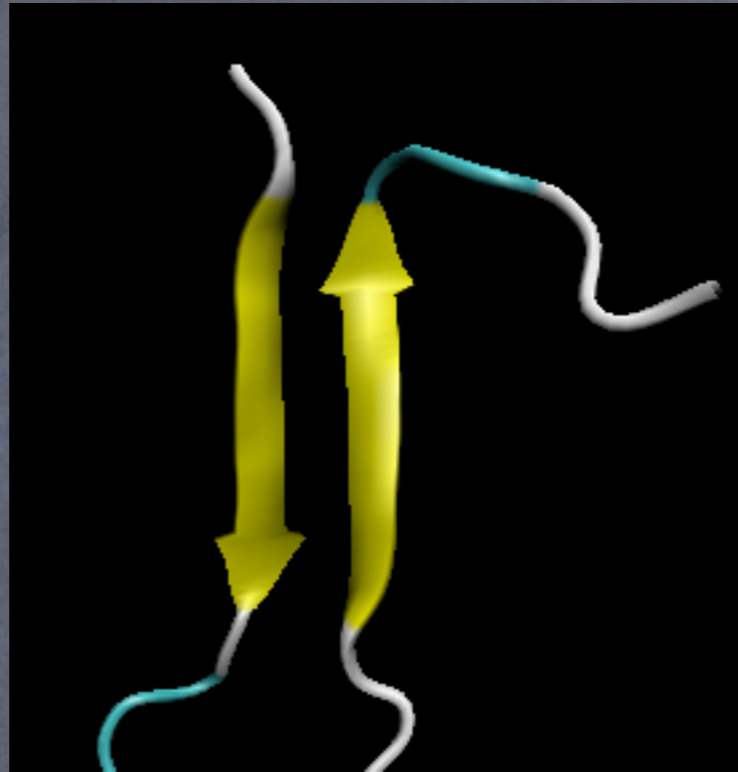
amino acid



Number N of conformations of a protein with 101 amino acids, 3 angles per bond:

$$N = 3^{100} \sim 10^{47}$$

Physical Constraints



SCOP database: 971 folds

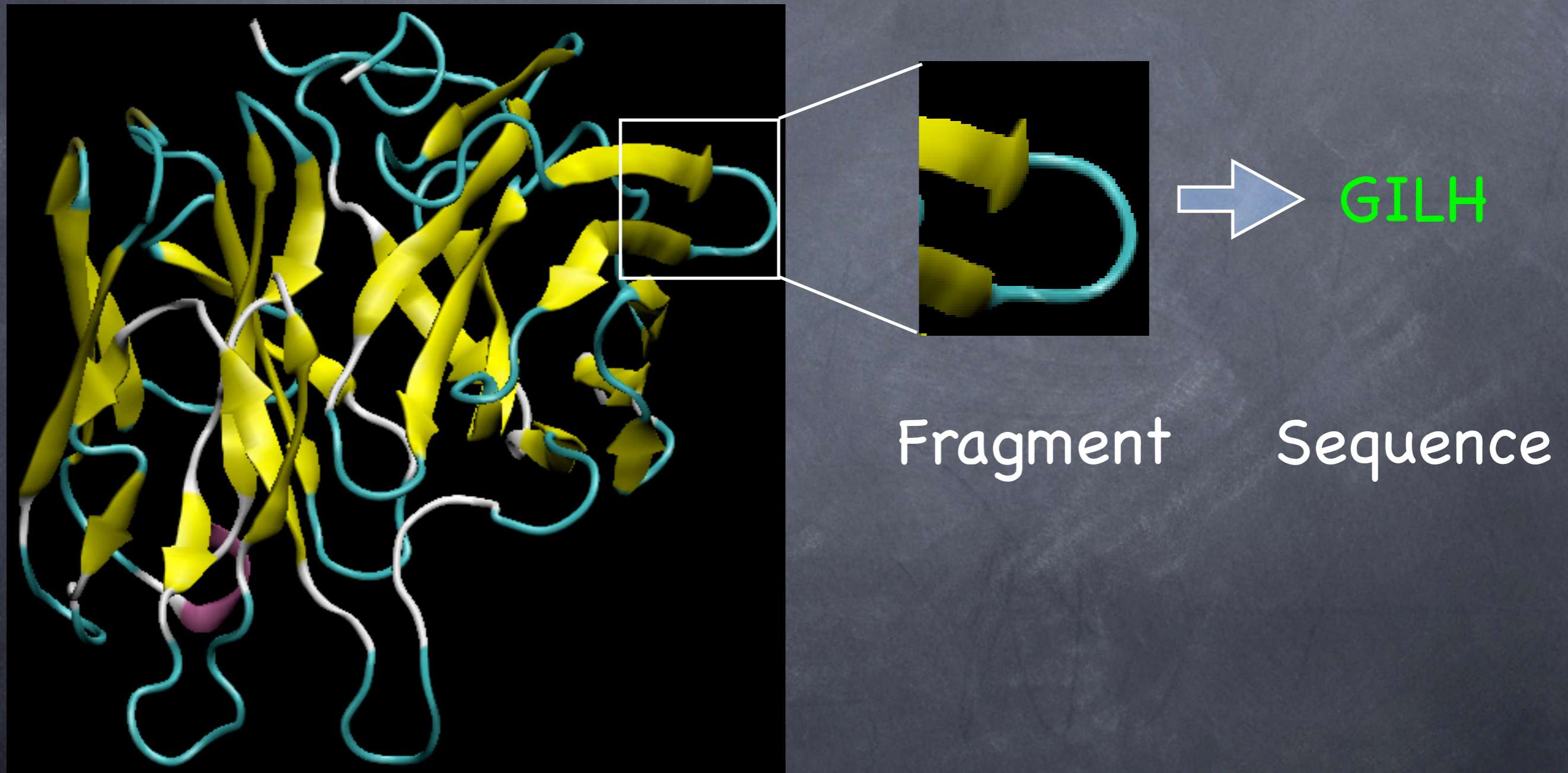
Estimated number of human genes: ca. 25000

Alexandrov & Go: predicted 16000 folds

Basic Questions

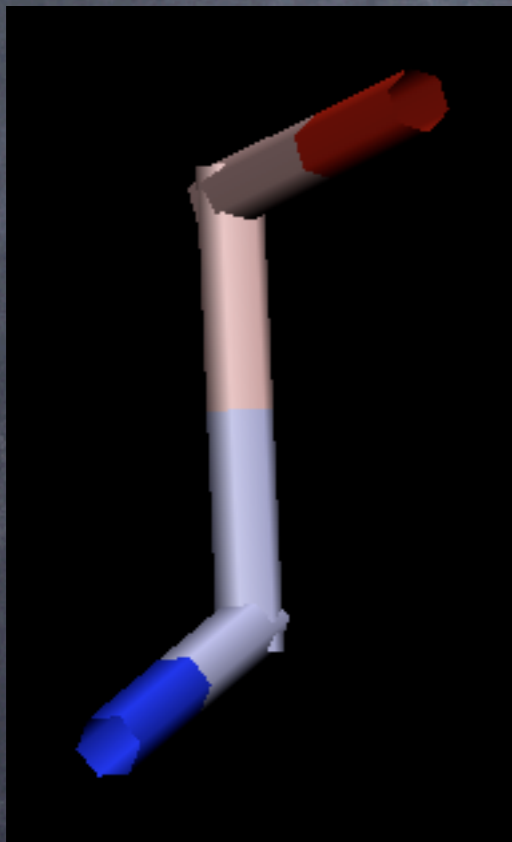
- How do we deal with all proteins simultaneously?
- Can we find a reduced set of the PDB?

MinSet: Proteomics Analysis on The Fragment Level

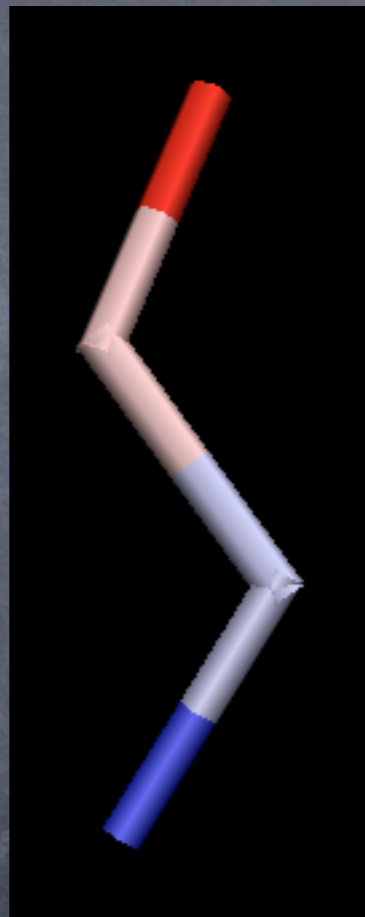


Representative Oligons

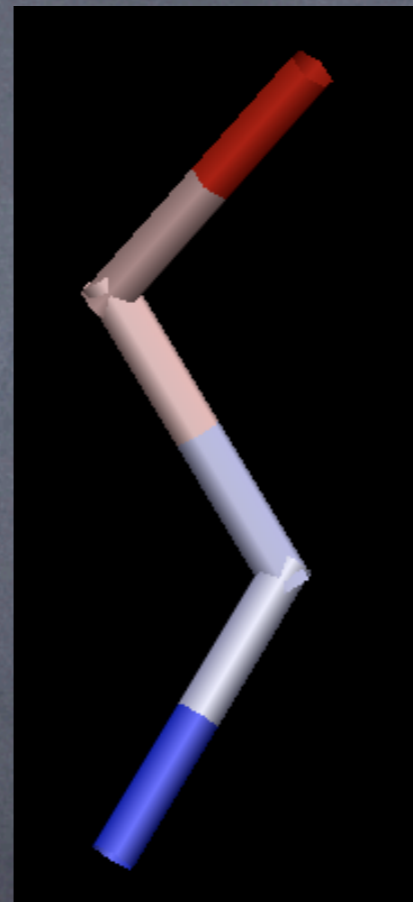
Only 28 different prototype fragments
of length $k=4$ (residues)!



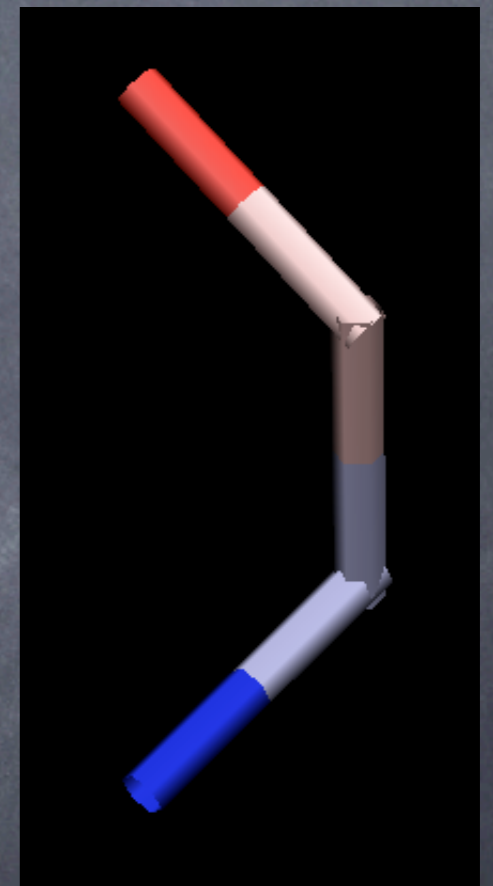
oligon 1



oligon 2



oligon 3



oligon 4

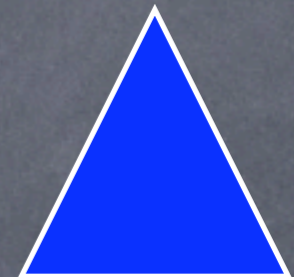
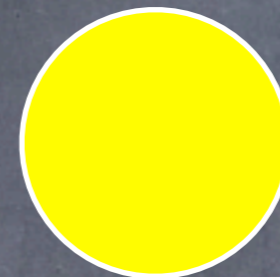
SA-27: A Structural Alphabet

- Definition of structural fragments $k=4$
- Overlapping fragments \Rightarrow structure
- Hidden Markov Model of protein structures
- Optimal set of fragments (Bayes. Inf. Crit.)
- 27 fragments expressed as alphabet

The MinSet Scheme

Structures

PDB



Structural Sequences

SA-27

agklew

yqweop

laqxcvq

dfnmxo

Genetic Algorithm

Subset

0

1

1

0

MinSet (1): Suffix Tree

SA-27

agklew

yqweop

laqxcvq

dfnmxo



Structure
Sequence

agklew-yqweop-laqxcvq-dfnmxo



Suffix Tree



Fragment Statistics

Basic Questions

- ✓ We can deal with all proteins simultaneously!
- Can we find a reduced set of the PDB?


MinSet (2): Genetic Algorithm

- Population size: 2500
- Generations: max. 100
- Top 10% selected
- Crossover, elitism
- Fitness: Fragment (k-word) entropy

Fitness Score and Coverage

Shannon Entropy

fragment
probability



$$H = - \text{Sum} [p \log(p)]$$

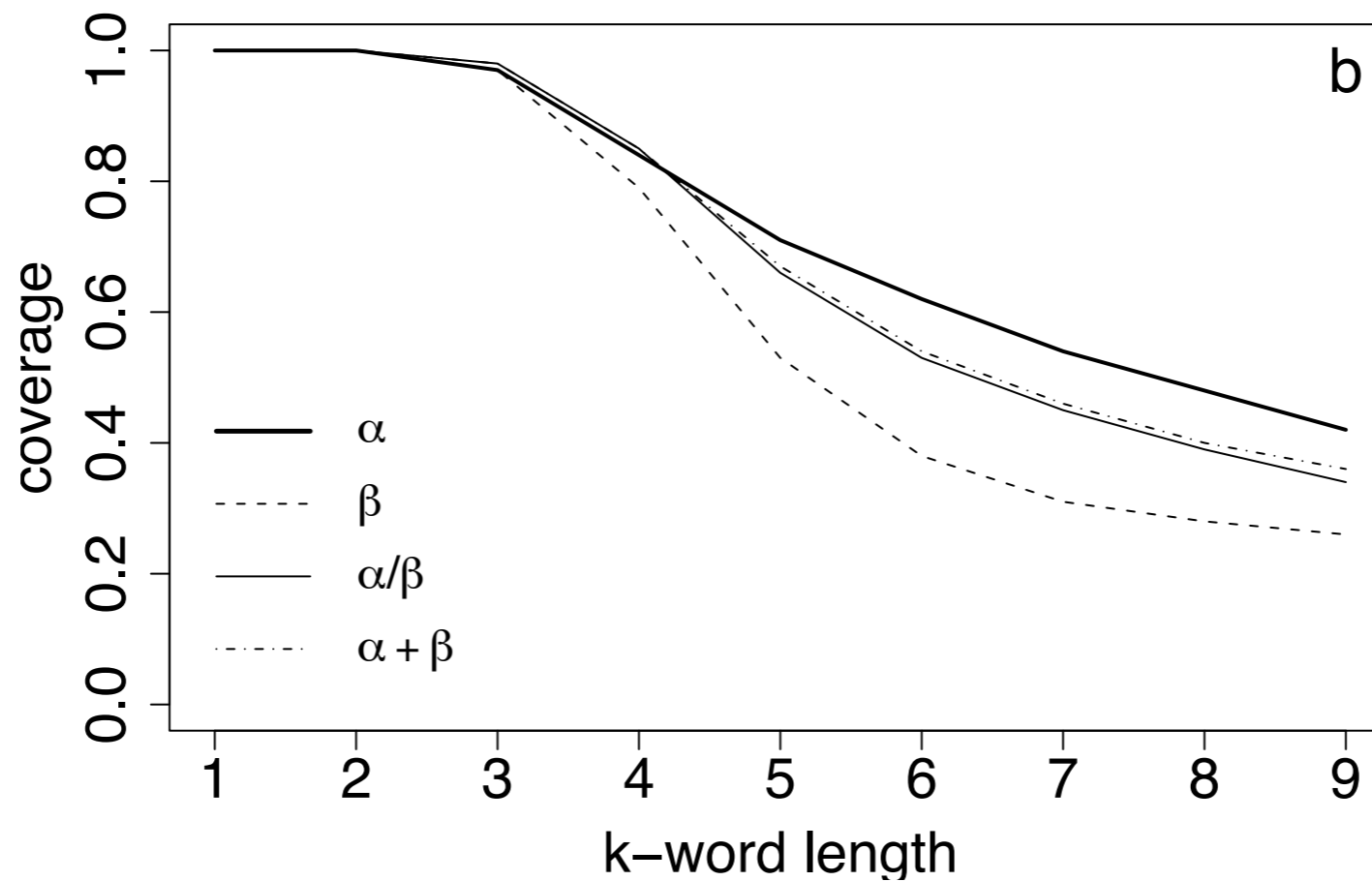
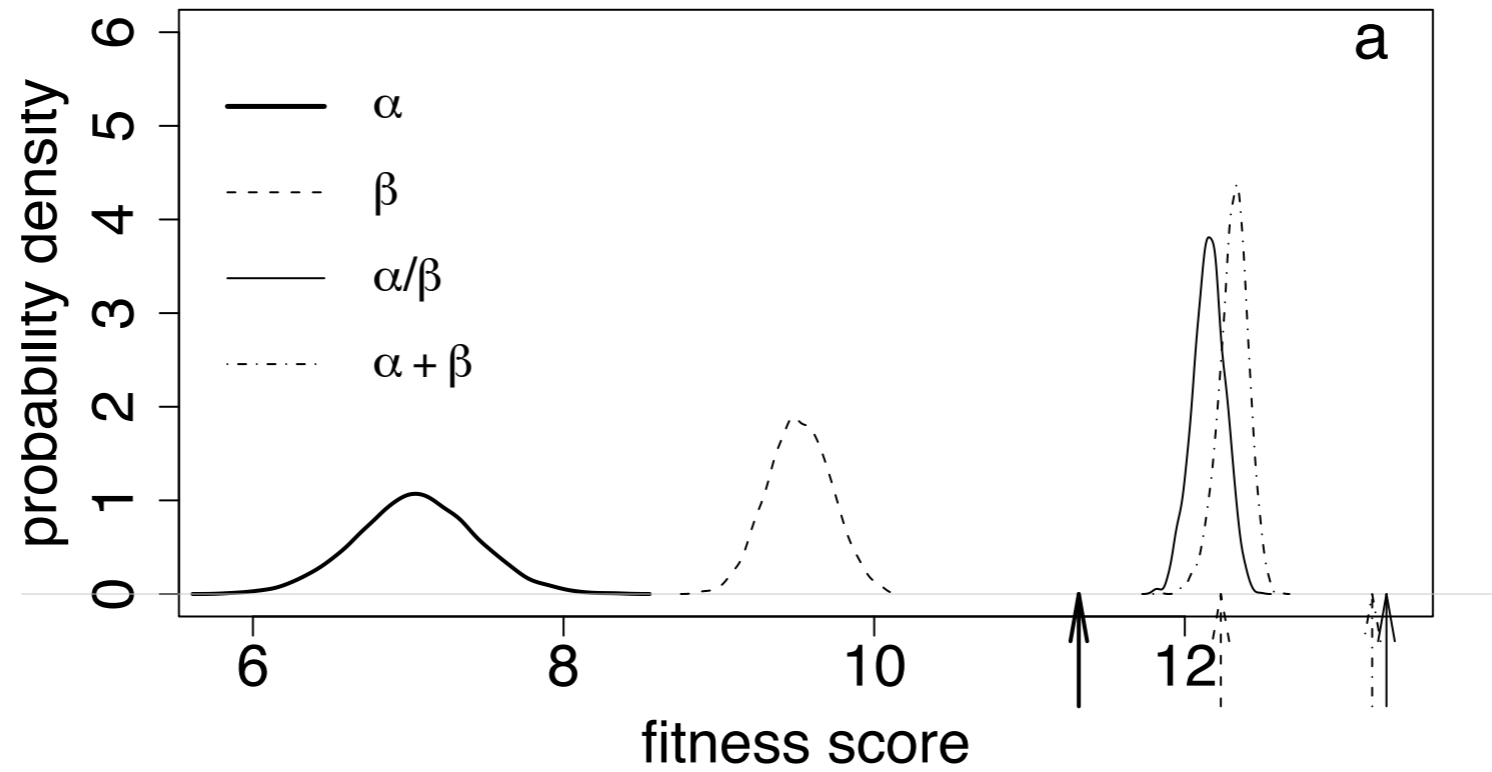
Kullback-Leibler Divergence

$$D = \text{Sum} [p \log(p/q)]$$

Fitness Score

$$S = H (1 - D)$$

MinSet of SCOP database



Basic Questions

- ✓ We can deal with all proteins simultaneously!
- ✓ Can we find a reduced set of the PDB?

MinSet Web Server



Upload a list of SCOP40 protein domains
(base set):

Alphabet:

Subset Size (%):

k-Word Length:

Provide job name (optional):

[Help on Options](#)

MinSet Database Online

Subsets

The data below are subsets of the SCOP40 (v1.69) database extracted from the basesets with the MinSet parameters 'selection k-word length' $k_s = 3, 5$ and 7 and 'target size' $t = 5\%, 10\%, 15\%$ and 20% .

Subsets with Target Size $t = 5\%$

Statistics

k_s	class	domain list	structure sequence	included k-words	excluded k-words
3	α	l3p05_a	l3p05_a.seq	inc.a.3.05.list	exc.a.3.05.list
3	β	l3p05_b	l3p05_b.seq	inc.b.3.05.list	exc.b.3.05.list
3	α/β	l3p05_c	l3p05_c.seq	inc.c.3.05.list	exc.c.3.05.list
3	$\alpha+\beta$	l3p05_d	l3p05_d.seq	inc.d.3.05.list	exc.d.3.05.list
5	α	l5p05_a	l5p05_a.seq	inc.a.5.05.list	exc.a.5.05.list
5	β	l5p05_b	l5p05_b.seq	inc.b.5.05.list	exc.b.5.05.list
5	α/β	l5p05_c	l5p05_c.seq	inc.c.5.05.list	exc.c.5.05.list
5	$\alpha+\beta$	l5p05_d	l5p05_d.seq	inc.d.5.05.list	exc.d.5.05.list
7	α	l7p05_a	l7p05_a.seq	inc.a.7.05.list	exc.a.7.05.list
7	β	l7p05_b	l7p05_b.seq	inc.b.7.05.list	exc.b.7.05.list
7	α/β	l7p05_c	l7p05_c.seq	inc.c.7.05.list	exc.c.7.05.list
7	$\alpha+\beta$	l7p05_d	l7p05_d.seq	inc.d.7.05.list	exc.d.7.05.list

Acknowledgements

Alessandro Pandini, Milano

Laura Bonati, Milano

Franca Fraternali, London