Bag-of-phrases Representation for Proteins and Proteomes

Pavle Goldstein and Braslav Rabar,
with JD and MZ

Department of Mathematics, University of Zagreb, Croatia

FOI2017

problem:

classify proteins in following organisms:

- *Streptomyces avermitilis*
- *Streptomyces coelicolor*
- and maybe in:
- *Arabidopsis thaliana*

really?

- classify by what criteria? origin, structure, function?
- do we know all the classes? (hint: no)
- should classification be universal? (hint: yes)

background on proteins:

- proteins are finite sequences in amino acid alphabet (length=20)
- sequence (=primary structure) determines tertiary structure and function
- structure prediction
- function prediction

background:

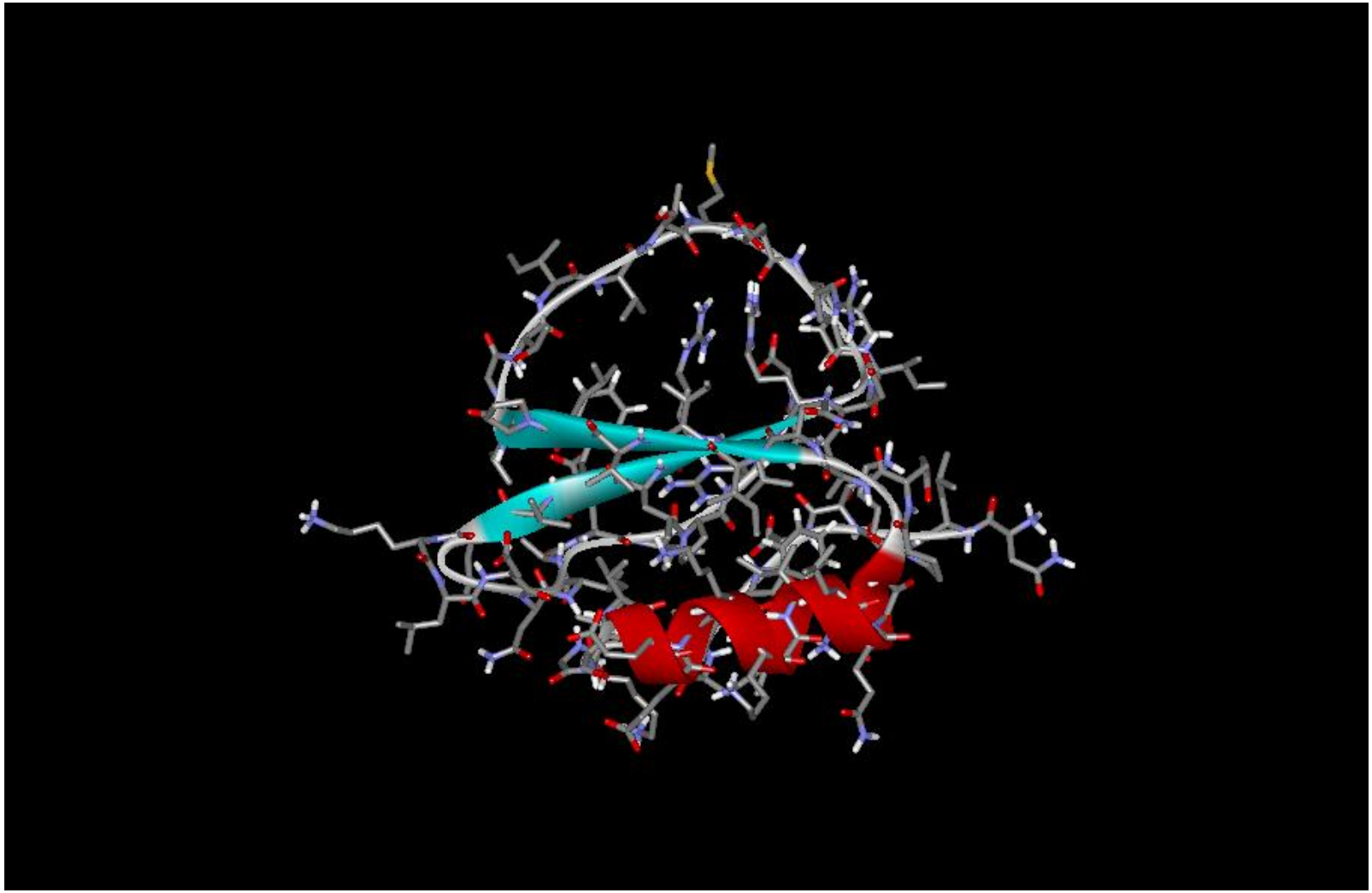

Fig : a protein structure model

background:

- protein primary structure:
- . . . MTNVTGDYTDCTPLLGDRAALDSFYEEHGYL. . .

solution, so far:

- plenty of databases, classify proteins with respect to various properties (e.g. Pfam. . .)
- not complete, new sequences arriving daily

bag-of-words representation

- documents, terms . . .
- frequencies, document-term matrix
- can it be applied to proteins?

b-o-w, continued

- documents $\sim$ proteins
- terms $\sim$????
- problem: no dictionary

b-o-w, still

- models with di-grams or tri-grams considered
- very noisy
- to work, needs serious training

dictionary

- all-against-all comparison for a fixed length
- iterative refinement
- when is a result relevant?

dictionary, continued

for example:

MALAGALA

MMMMGALA

AAAAGALA

MALALLLL

MALAGGGG

dictionary, continued

- need criteria for biological relevance
- maximal clique, communities in graphs
- eigenvalues

PSSM assignment

- PSSM: position-specific scoring matrix, position weight matrix (PWM), or position-specific weight matrix (PSWM)
- PSSM for k columns: $y = y_1..y_k$
- sequence of length n: $x = x_1..x_n$
- marginal/background distribution of AA: $q = q_1..q_{20}$
- window-sliding method: evaluate window at positions $l, l+1, .., l+k-1$ by

$$s(l) = \sum_{i=0}^{k-1} \log \frac{P(x_{l+i}|y_{i+1})}{P(x_{l+i}|q)}$$

- PSSM-assignment: k-window $x_j, \ldots, x_{j+k-1}$ that maximizes the $s(l)$
- log-likelihood ratio statistic

b-o-p

- document (=proteome) contains approx. 8500 documents (=proteins), av. length 400
- we got approx. 20 000 "relevant" phrases (length 10)
- further trimming and merging - approx 8000 words with profiles
- hence, approx. 8000 models, length 10-30

b-o-p, still

- warning: profiles are variable
- not a single language
- more like a family
- how to measure frequency? alternatives?

b-o-p representation

- we measure the length of the best match
- $s(l) = \sum_{i=0}^{k-1} \log \frac{P(x_{l+i}|y_{i+1})}{P(x_{l+i}|q)}$
- asymptotically equivalent

comments

- important phrases recognizable
- construction agrees with protein evolution
- drawback: a single best match

results

- example: ATP-binding "family" from *Streptomyces coelicolor*
- approx. 140 proteins, great variability
- support reduction
- not close in euclidean norm