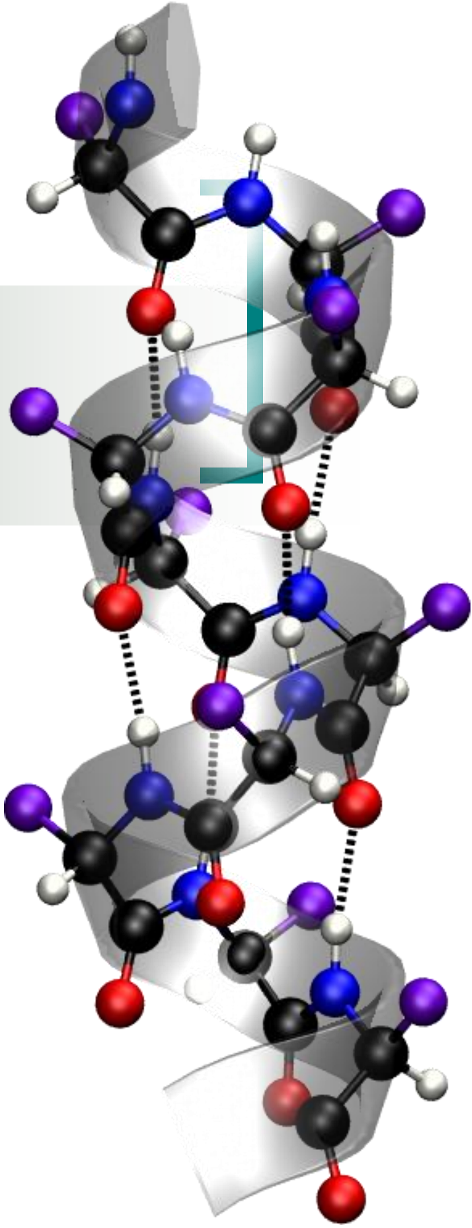
Similarity Search in Protein Databases



David Hoksza

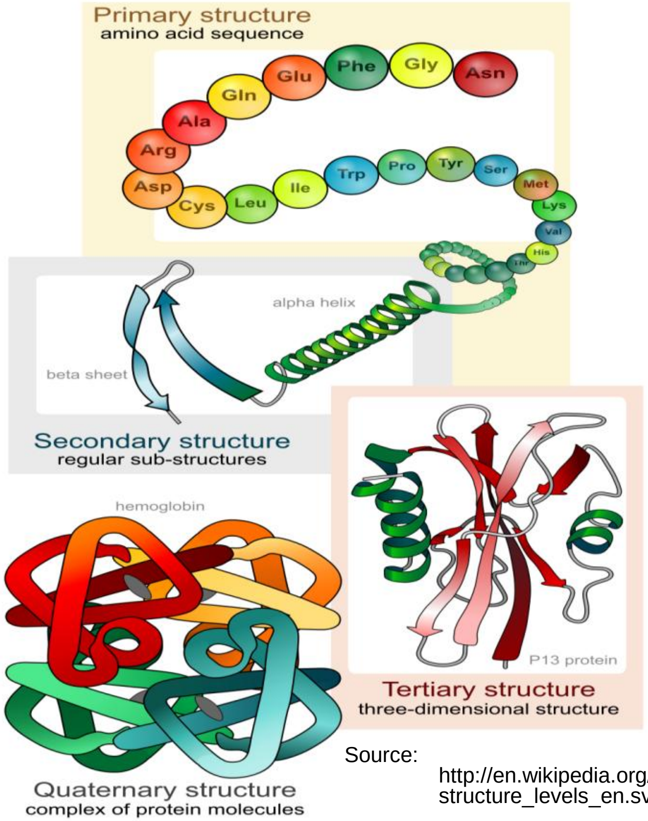
hoksza@ksi.mff.cuni.cz

Thesis Contributions

- Protein sequence similarity
  - metric indexing approach
  - sequential approach
  - speedup within existing similarity model
- Protein structure similarity
  - nonalignment-based similarity approach
  - alignment-based similarity approach
  - similarity modelling itself

Protein structure & Motivation

- Transportation, building, signalling, catabolism, ...
- Molecule consisting of 20 types of amino acids (AA)



Source: http://en.wikipedia.org/wiki/File:Main_protein_structure_levels_en.svg

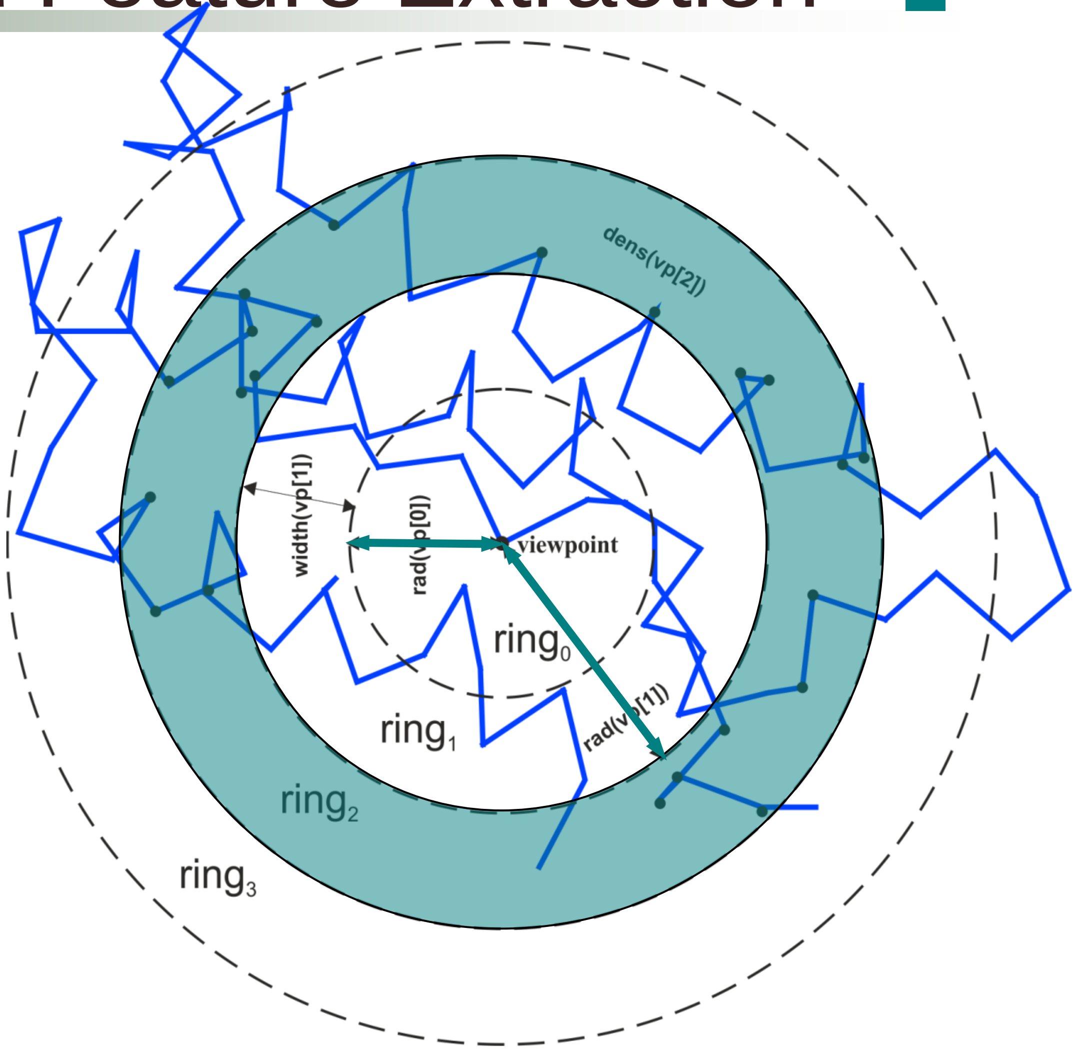
- Central dogma of molecular biology
  - DNA → RNA → protein
- Proteins' 3D interactions secure biological function
- **protein structure similarity → biological function similarity**
- **protein sequence similarity → protein structure similarity**

Protein Structure Similarity

- Nonalignment-based approach
  - indexing
  - feature extraction
  - simple or ad-hoc similarity measure
- Alignment-based approach
  - sequential scan
  - feature extraction
  - alignment
  - superposition (3D transformation)
  - similarity measure
    - RMSD, TM-score
- Quality criterion
  - classification accuracy (SCOP hierarchical classification)

Density-Based Feature Extraction

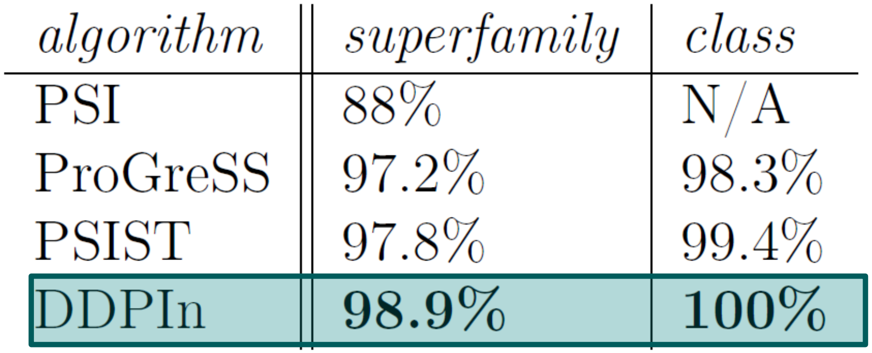
- Features
  - n-dimensional vectors of real numbers
  - AA ≈ viewpoint → VPT (viewpoint tag)
- sDens
  - density of AAs in rings of predefined width
- sRad
  - widths of rings containing predefined percentage of AAs



D. Hoksza. **DDPIn: Distance and Density-Based Protein Indexing**, CIBCB 2009, **IEEE**

Nonalignment-Based Approach

- One-step search
  - database creation
    1. AAs → feature vectors
    2. indexing using weighted L_2 metric and MAM
  - querying
    1. AAs → feature vectors
    2. feature vector → query object
    3. results’ merging
    4. SCOP classification

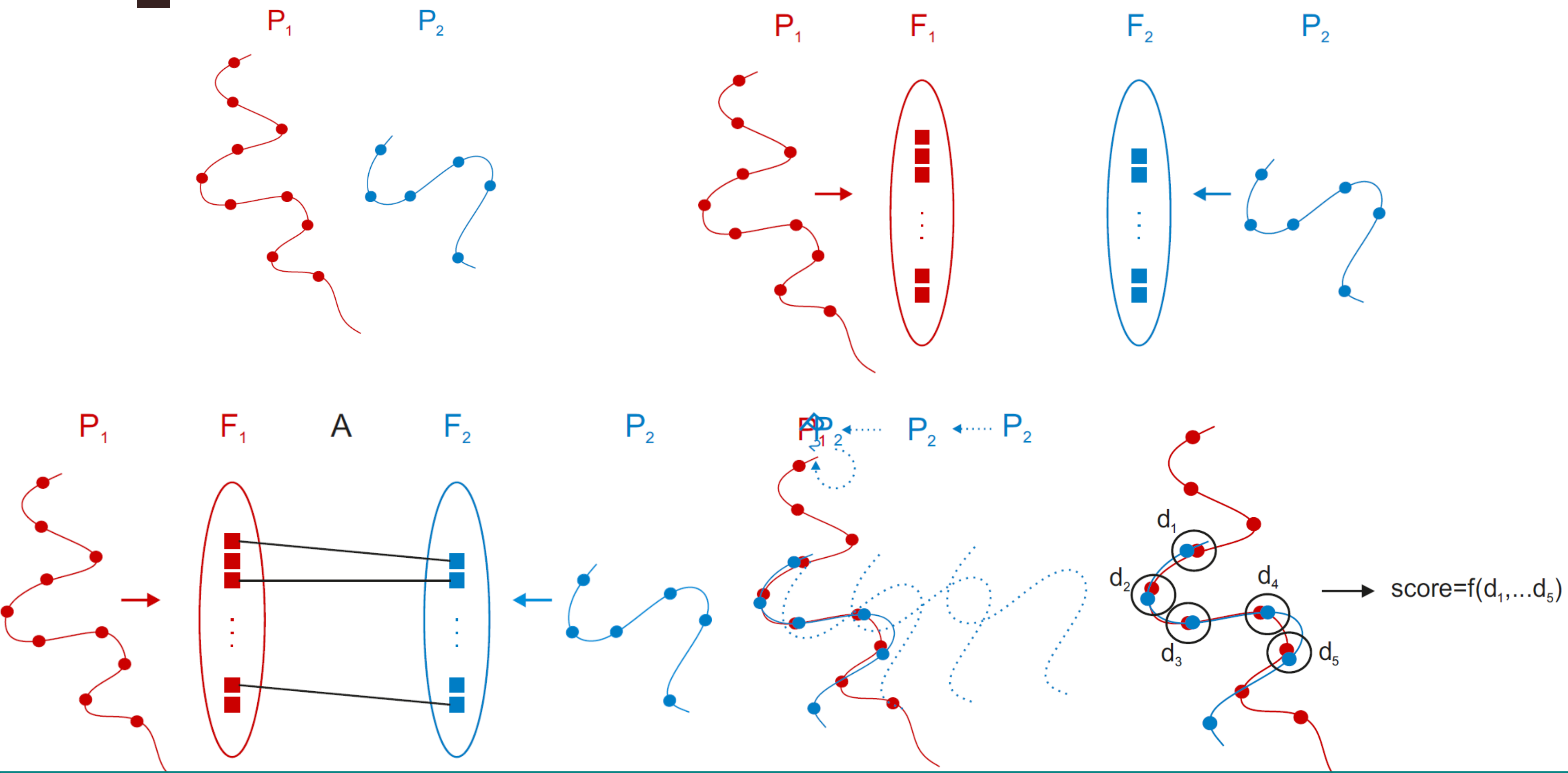


algorithm	*superfamily*	*class*
PSI	88%	N/A
ProGreSS	97.2%	98.3%
PSIST	97.8%	99.4%
DDPIn	**98.9%**	**100%**

- Two-step search
  - 2 one-step searches
  - results’ comparison
  - rescoring using Smith-Waterman
  - SCOP classification

D. Hoksza. **DDPIn: Distance and Density-Based Protein Indexing**, CIBCB 2009, **IEEE**

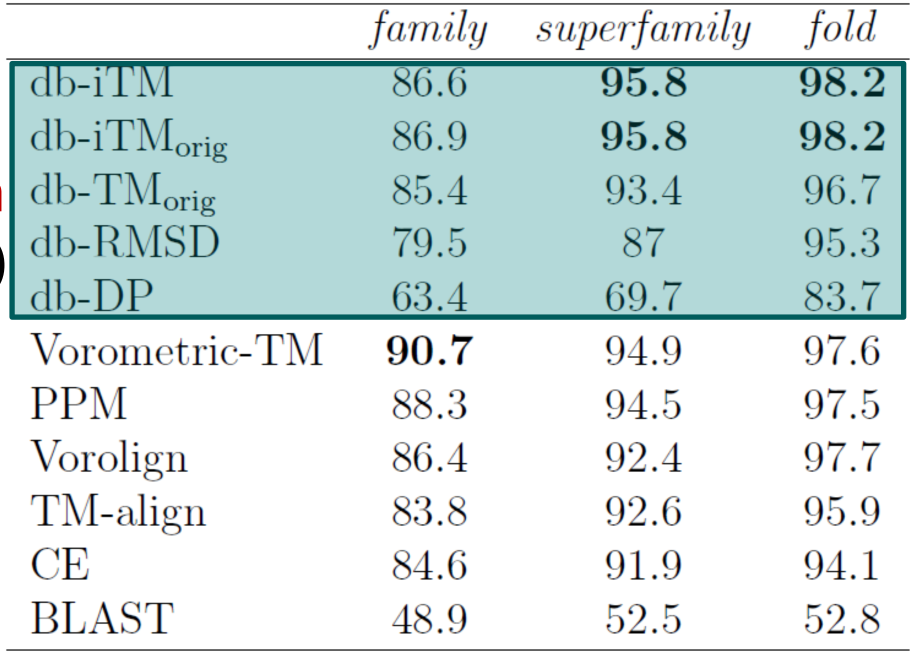
Alignment-Based Approach



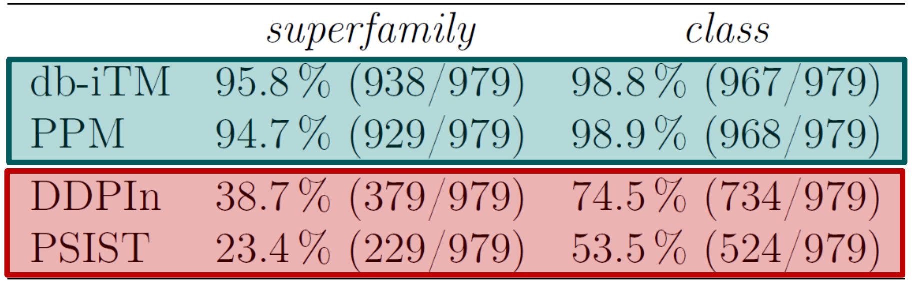
D. Hoksza, J. Galgonek. **Density-Based Classification of Protein Structures Using Iterative TM-score**, BIBMW 2009, **IEEE**
D. Hoksza, J. Galgonek. **Alignment-Based Extension to DDPIn Feature Extraction**, IJCB, **ACTA Press**, 2010

Alignment-Based Approach

- Feature extraction
  - AAs → feature vectors
  - density-based feature extraction
- Alignment (amino acid matching)
  - Smith-Waterman alignment
    - distance between feature vectors → scoring matrix
    - modified variable gap penalty system
- Superposition + scoring
  - RMSD
  - TM-score
    - reducing number of initial states
    - iterative dynamic programming with belt-based restriction



	family	*superfamily*	*fold*
db-iTM	86.6	**95.8**	**98.2**
db-iTM$_{orig}$	86.9	**95.8**	**98.2**
db-TM$_{orig}$	85.4	93.4	96.7
db-RMSD	79.5	87	95.3
db-DP	63.4	69.7	83.7
Vorometric-TM	**90.7**	94.9	97.6
PPM	88.3	94.5	97.5
Vorolign	86.4	92.4	97.7
TM-align	83.8	92.6	95.9
CE	84.6	91.9	94.1
BLAST	48.9	52.5	52.8



	superfamily	*class*
db-iTM	95.8 % (938/979)	98.8 % (967/979)
PPM	94.7 % (929/979)	98.9 % (968/979)
DDPIn	38.7 % (379/979)	74.5 % (734/979)
PSIST	23.4 % (229/979)	53.5 % (524/979)

D. Hoksza, J. Galgonek. **Density-Based Classification of Protein Structures Using Iterative TM-score**, BIBMW 2009, **IEEE**

D. Hoksza, J. Galgonek. **Alignment-Based Extension to DDPIn Feature Extraction**, IJCB, **ACTA Press**, 2010

Alignment-Based Approach - Indexing

- Indexability measures
  - intrinsic dimensionality
    - objects' distribution
  - ball overlap factor (BOF)
    - ball regions' separation
  - T-error
    - nontriangular triplets

- Semimetrization
  - $\text{db-iTM}_\mathfrak{f} = 1 - \text{db-iTM}$
  - $\text{db-iTM}_\text{smf}(S_i, S_j) = max(\text{db-iTM}_\mathfrak{f}(S_i, S_j), \text{db-iTM}_\mathfrak{f}(S_j, S_i))$
  - reranking

- Metrization
  - Trigen

measure	*intrinsic dimensionality*	*T-error*	*BOF*
db-iTM_smf	131.2	0.000005 %	96.8 %
$\text{db-iTM}_\text{smf}^{2.5}$	24.3	0.04 %	58.1 %
$\text{db-iTM}_\text{smf}^{3}$	17.5	0.10 %	44.5 %
$\text{-log}(1\text{-db-iTM}_\text{smf})$	6.9	0.15 %	44.4 %

J. Galgonek, D. Hoksza. **On the Effectiveness of Distances Measuring Protein Structure Similarity**, SISAP 2009, **IEEE**

Conclusion

- Protein sequence search
  - **indexing**
    - data domain exploration
  - **speeding distance computation**
    - 20% speedup
    - expected speedup growth
- Protein structure search
  - **nonalignment-based**
    - best accuracy on all SCOP levels
  - **alignment-based**
    - best accuracy on superfamily and fold SCOP levels
    - indexing possibilities
    - comparison with nonalignment-based

Feature Work

- Protein structure similarity
  - indexing
  - application of biological features
- RNA secondary structure similarity
- RNA tertiary structure similarity