The GAssist Pittsburgh Learning Classifier System

Dr. J. Bacardit, N. Krasnogor
G53BIO - Bioinformatics

asap research
automated
scheduling
optimisation
& planning

The University of Nottingham

Outline

- GAssist applied to bioinformatics
- Summary and future directions

Objectives of GAssist

- GAssist [Bacardit, 04] is a Pittsburgh Approach Learning Classifier System evolving variable-length rule sets
- The research done on this system has three objectives
  - Generation of compact and accurate solutions
  - Run-time reduction
  - Representations for real valued attributes

Objectives of GAssist

- Representations for real valued attributes
  - GAssist should be applicable to a range of problems as broad as possible
  - This means that it should be able to handle continuous attributes
  - Achieved by the The Adaptive Discretization Intervals (ADI) rule representation

GAssist applied to Bioinformatics

- GAssist has been applied to protein domains
- Proteins are biological molecules of primary importance to the functioning of living organisms
- Proteins are constructed as a chains of amino acid residues
- This chain folds to create a 3D structure

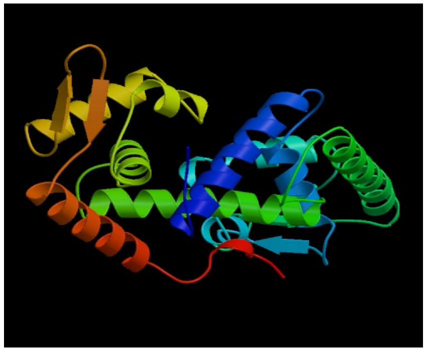
GAssist applied to Bioinformatics

- It is relatively easy to know the primary sequence of a protein, but much more difficult to know its 3D structure
- Can we predict the 3D structure of a protein from its primary sequence? → Protein Structure Prediction (PSP)
- PSP problem is divided in several sub problems. We focus on **Coordination Number (CN) prediction**

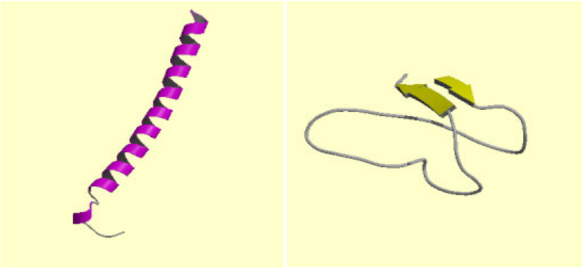
GAssist applied to Bioinformatics

Primary Structure = Sequence

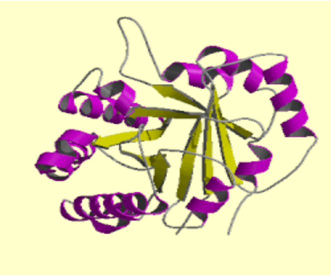
```
MKYNNHDKIRDFIIIEAYMFRFKKKVKPEVDMTIKEFILLTY
LFHQQENTLPFKKIVSDLCYKQSDLVQHIKVLVKHSYISKV
RSKIDERNTYISISEEQREKIAERVTLFDQIIKQFNLADQSE
SQMIPKDSKEFLNLMMYTMYFKNIIKKHLTLSFVEFTILAIIT
SQNKNIVLLKDLIETIHHKYPQTVRALNNLKKQGYLIKERS
TEDERKILIHMDDAQQDHAEQLLAQVNQLLADKDHLHLVF
E
```



Secondary Structure



Tertiary

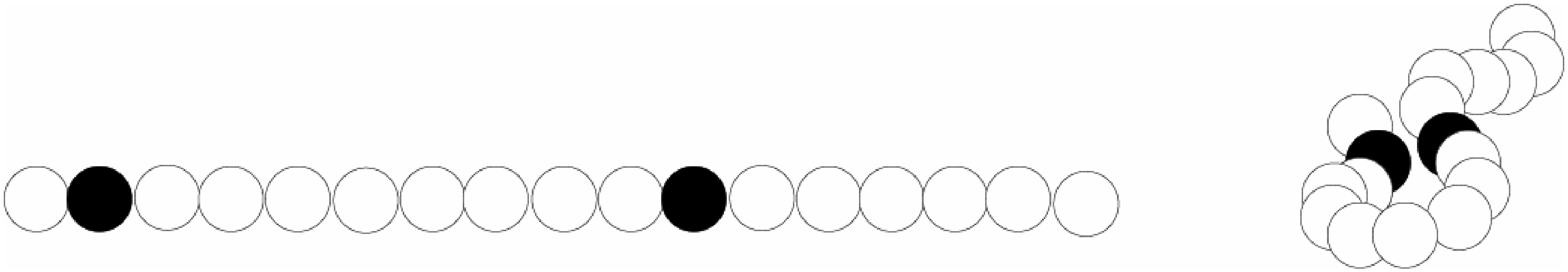


Local Interactions

Global Interactions

GAssist applied to Bioinformatics

- Coordination Number (CN) prediction
  - Two residues of a chain are said to be in contact if their distance is less than a certain threshold
  - **CN of a residue :** number of contacts that a certain residue has



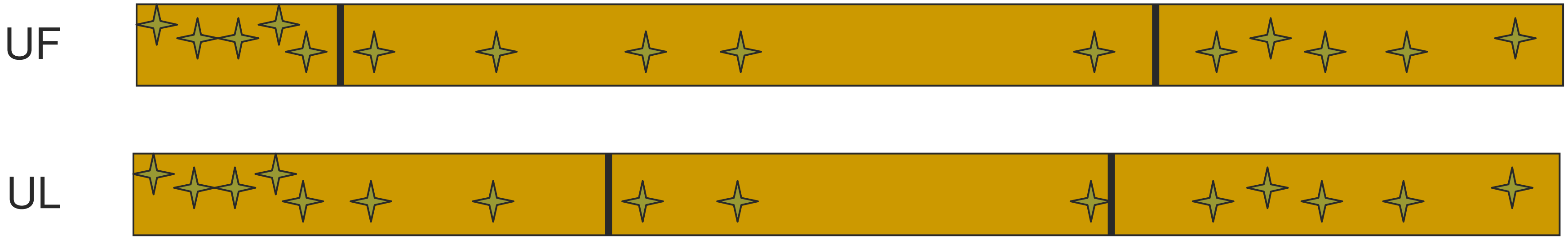
GAssist applied to Bioinformatics

- Kinjo et al.'s definition of CN
  - Distance between two residues is defined as the distance between their C_β atoms (C_α for Glycine)
  - Uses a smooth definition of contact based on a sigmoid function instead of the usual crisp definition
  - Discards local contacts

$$O_i^p = \sum_{j:|j-i|>2} \frac{1}{1+exp(w(r_{ij}-d_c))}$$

GAssist applied to Bioinformatics

- Classification approach
  - We need to convert the real-valued CN into a finite set of categories
  - We have tested two criteria based on the two usual unsupervised discretization methods: Uniform Frequency (UF) and Uniform Length (UL)

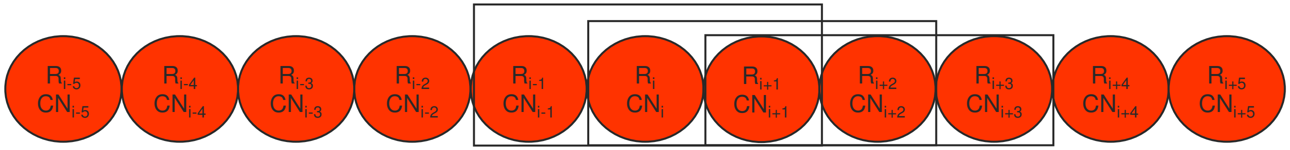


GAssist applied to Bioinformatics

- Protein dataset
  - Used the same set used by Kinjo et al.
  - 1050 protein chains
  - 259768 residues
  - Ten lists of the chains are available, first 950 chains in each list are for training, the rest for tests (10xbootstrap)

GAssist applied to Bioinformatics

- We have to transform the data into a regular structure so that it can be processed by standard machine learning techniques
- Each residue is characterized by several features. We use one (i.e., the AA type) or more of them as input information and one of them as target (CN)



$R_{i-1},R_i,R_{i+1} \rightarrow CN_i$
$R_i,R_{i+1},R_{i+2} \rightarrow CN_{i+1}$
$R_{i+1},R_{i+2},R_{i+3} \rightarrow CN_{i+2}$

GAssist applied to Bioinformatics

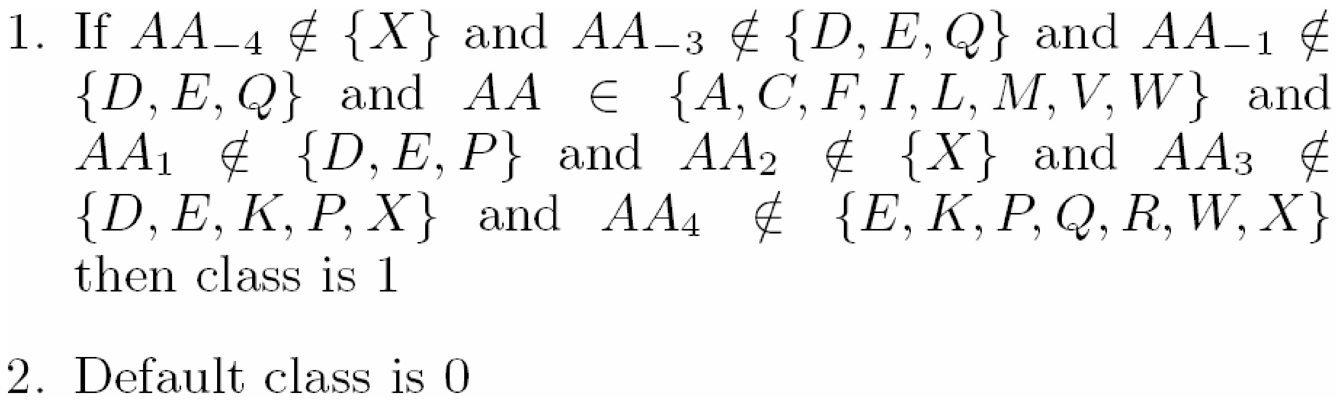
- Input information
  - 3 types of input information
    - Base information: The AA type of the residues included in the window around the target
    - Global protein information
      - Aim: providing information about the average CN of the protein chain
      - 1st version: 21 attributes: length of the protein and frequency of appearance of the 20 AA types
      - 2nd version: 1 attribute: predicted ave. CN using the 21 att. defined above as input
    - Predicted SS of the residues included in the window

GAssist applied to Bioinformatics

- Summary of results
  - All datasets using UL class definition have better performance than their UF equivalent (7-12% dif.)
  - PredSS gives a 2-3% performance boost
  - Global protein information gives a 1.5-2% performance boost

GAssist applied to Bioinformatics

- Interpretability analysis of GAssist
  - Example of a rule set for the CN1-UL-2 classes dataset



1. If $AA_{-4} \notin \{X\}$ and $AA_{-3} \notin \{D, E, Q\}$ and $AA_{-1} \notin \{D, E, Q\}$ and $AA \in \{A, C, F, I, L, M, V, W\}$ and $AA_1 \notin \{D, E, P\}$ and $AA_2 \notin \{X\}$ and $AA_3 \notin \{D, E, K, P, X\}$ and $AA_4 \notin \{E, K, P, Q, R, W, X\}$ then class is 1
2. Default class is 0

  - All AA types associated to the central residue are hydrophobic (core of a protein)
  - D, E consistently do not appear in the predicates. They are negatively charges residues (surface of a protein)

GAssist applied to Bioinformatics

- Future directions
  - Assessing the added value of the class definitions
  - Testing other types of input information
  - Extending the interpretability analysis
  - Improving GAssist
    - With purely ML techniques
    - Feeding back information from the interpretability analysis to bias the search

Summary and future directions

- GAssist produces very compact but accurate rule sets
- This is done by combining the techniques described briefly in this presentation

Summary and future directions

- Future directions
  - Smart recombination operators
  - Develop theoretical models for all the components of the system