Models and Metaphors from Biology to Bioinformatics Tools - Camerino, Sept 5-7, 2004

A MultiAgent System for Protein Secondary Structure Prediction

G. Armano, G. Mancosu, and A. Orro

DIEE – University of Cagliari, Italy

email: {armano,mancosu,orro}@diee.unica.it

Outline of the Talk

- SSP at a Glance ...
- MASSP: a Software Architecture for SSP
- MASSP: Micro-Architecture
- Experimental Results
- Conclusions and Future Work

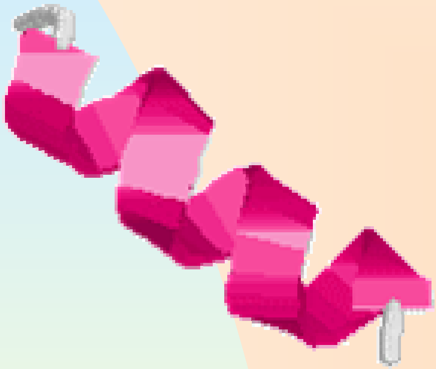
SSP at a Glance ...

Defining the problem of protein secondary structure prediction ...

The Problem in Hand ...

Myoglobin – J. Kendrew, 1960

The Problem in Hand ...



Alpha-helix

Beta-sheet

Protein SSP: Motivations

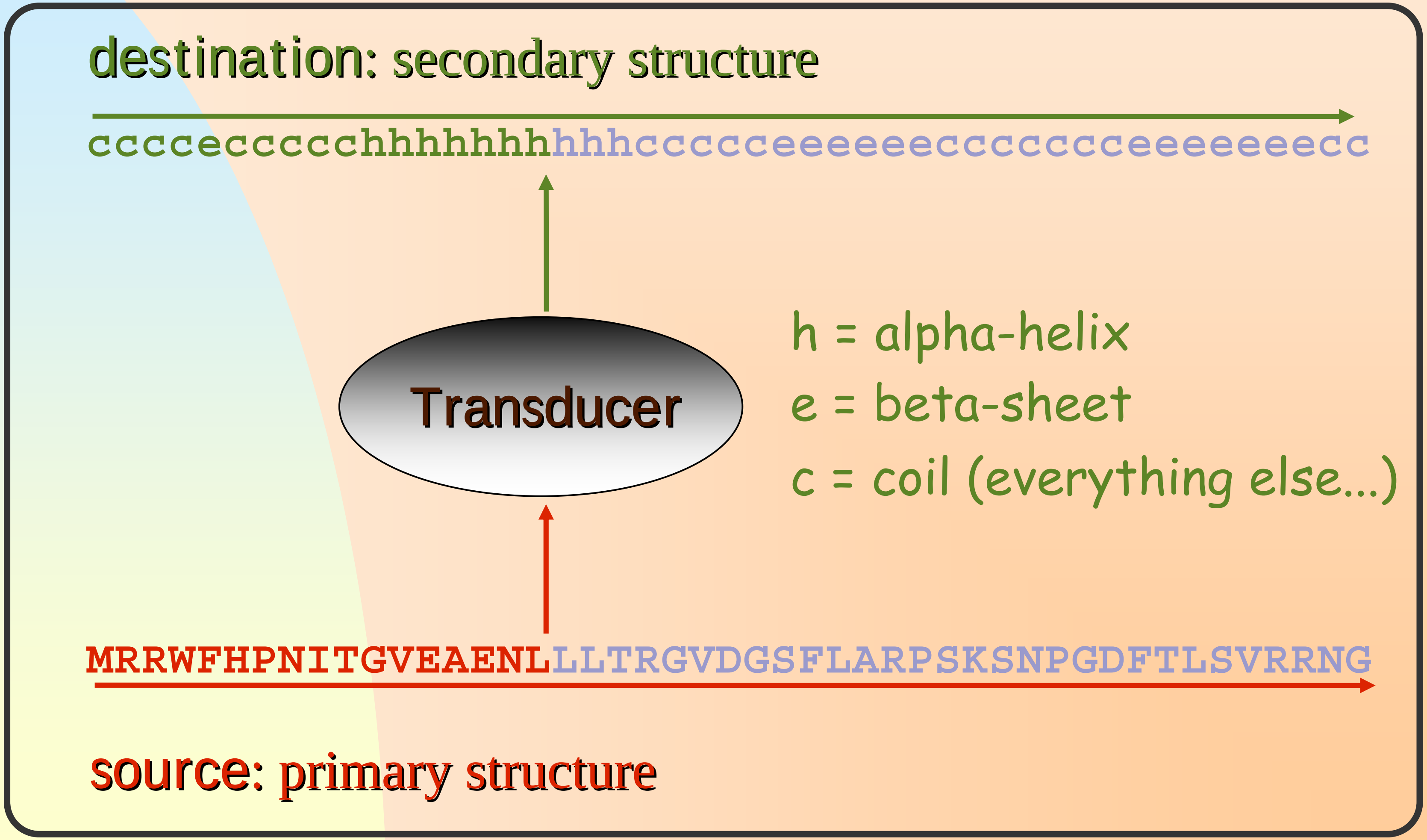
- There are lots of known tridimensional structures, determined by NMR and X-ray crystallography methods, and their number is rapidly growing
  - More than 25,000 proteins, in the PDB database, on June 2004
- On the other hand, the number of discovered proteins without a known structure is growing faster
  - 153,000 entries, in Swiss-Prot, on June 2004

Protein SSP: Motivations

- Predicting protein (3D) structure is a very complex task
- Most methodologies concentrate on the simplified task of predicting secondary structures
- The secondary structure of a protein can be useful to find information about its functionality (homology “through” similarity)

Protein SSP at a Glance ...

1ayab-1-GJB (CB396)



destination: secondary structure

cccceccccchhhhhhhhhhccccceeeeeecccccccceeeeeeecc

Transducer

h = alpha-helix

e = beta-sheet

c = coil (everything else...)

MRRWFHPNITGVEAENLLLTRGVDGSFLARPSKSNPGDFTLSVRRNG

source: primary structure

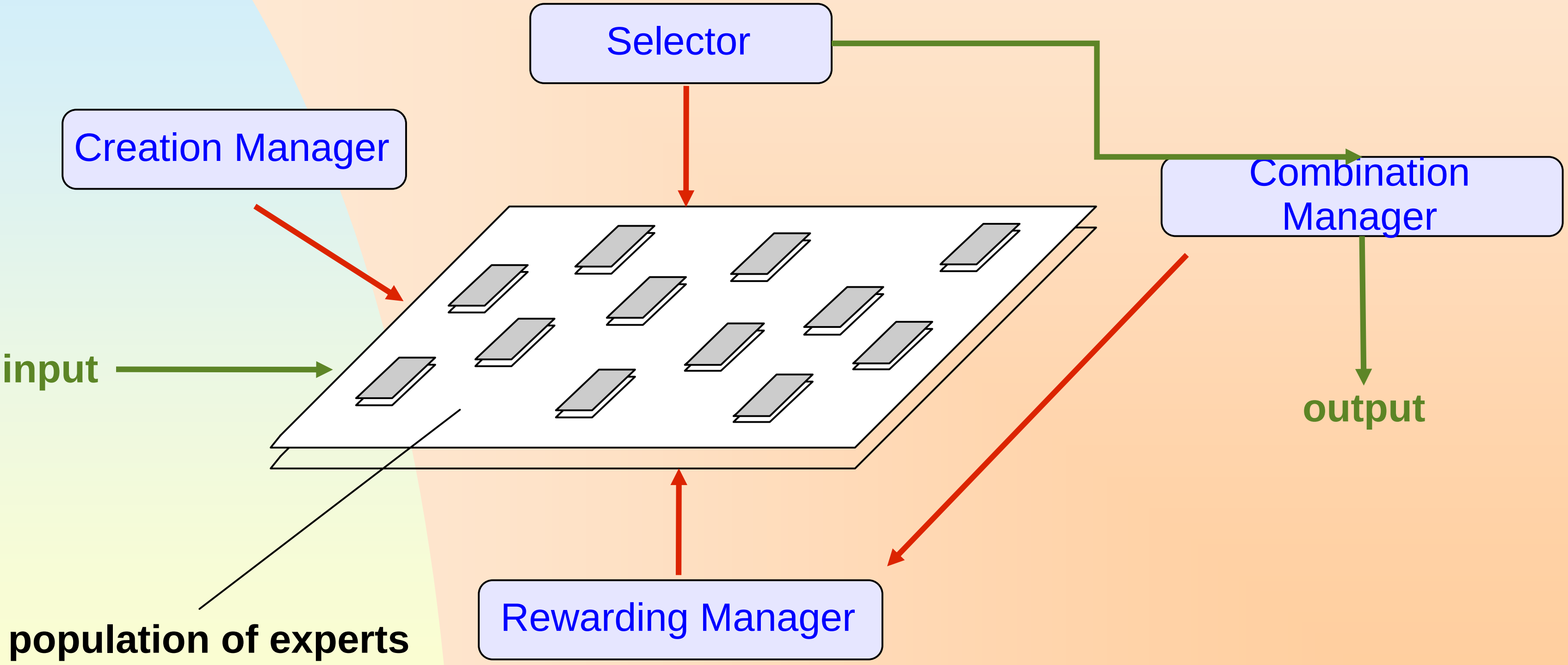
MASSP: A Software Architecture for SSP

Using multiple experts to predict protein secondary structure (*MASSP* = Multi-Agent SSP)

MASSP: Adopting Multiple Experts

- A population of experts has been used instead of a single expert for their capability of augmenting the overall accuracy, under the hypothesis of independence or negative correlation on errors

MASSP: Macro Architecture



MASSP: Most Relevant Features

- Offline / Online training strategy
- Hard / Soft region splitting, with overlapping
- Experts are locally scoped (an expert is able to deal with a –typically proper– subset of the inputs)
- Match-set formation (given an input, not all experts are involved in the prediction activity)
- Outputs combination throughout a weighted averaging
- Selective environment (at each epoch, experts can die or survive depending on their relative strength)

MASSP: Offline Strategy

- The training strategy is offline, meaning that a separate learning, validation, and test set are used (possibly averaging the results with N-fold cross-validation)

MASSP: Soft Region Splitting

- An expert can be more or less able to deal with a given input *x*
- The degree of expertise between an input and an expert is the result of a flexible matching activity and ranges over the interval [0,1]
- There is a many-to-many relation between experts and inputs

In particular, given an input, more that one expert is able to deal with it

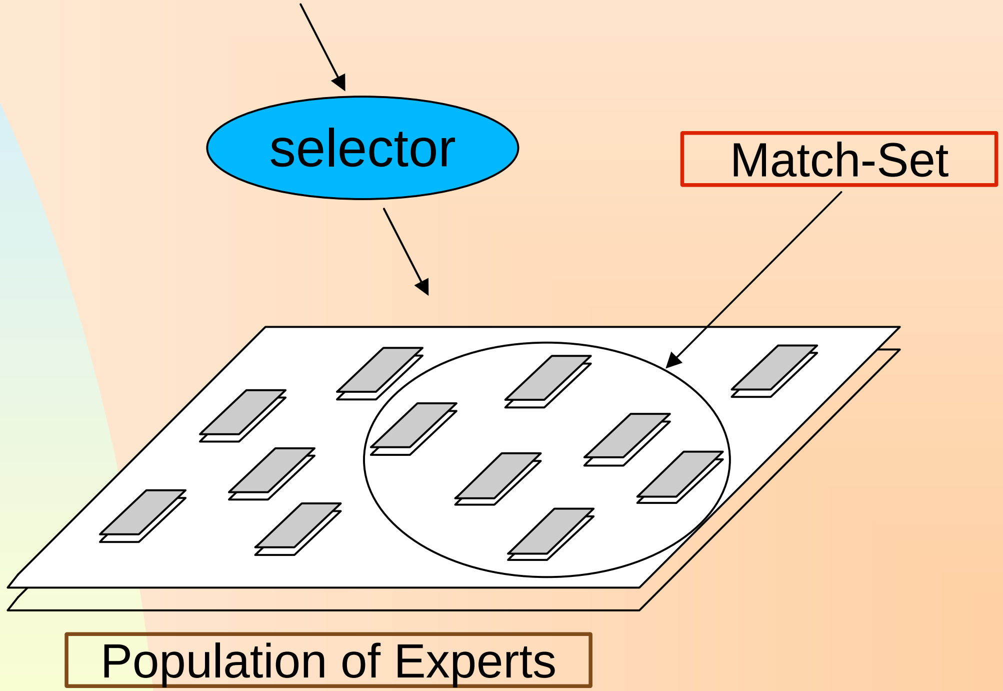
MASSP: Experts' Scope

- To make the learning task easier, each expert does not have a complete visibility of the input space

  In particular, given an input, several experts can be involved in predicting it –thus forming the match-set

MASSP: Match-Set Formation

Input: ... `msgkmtgivkwfnad` ...

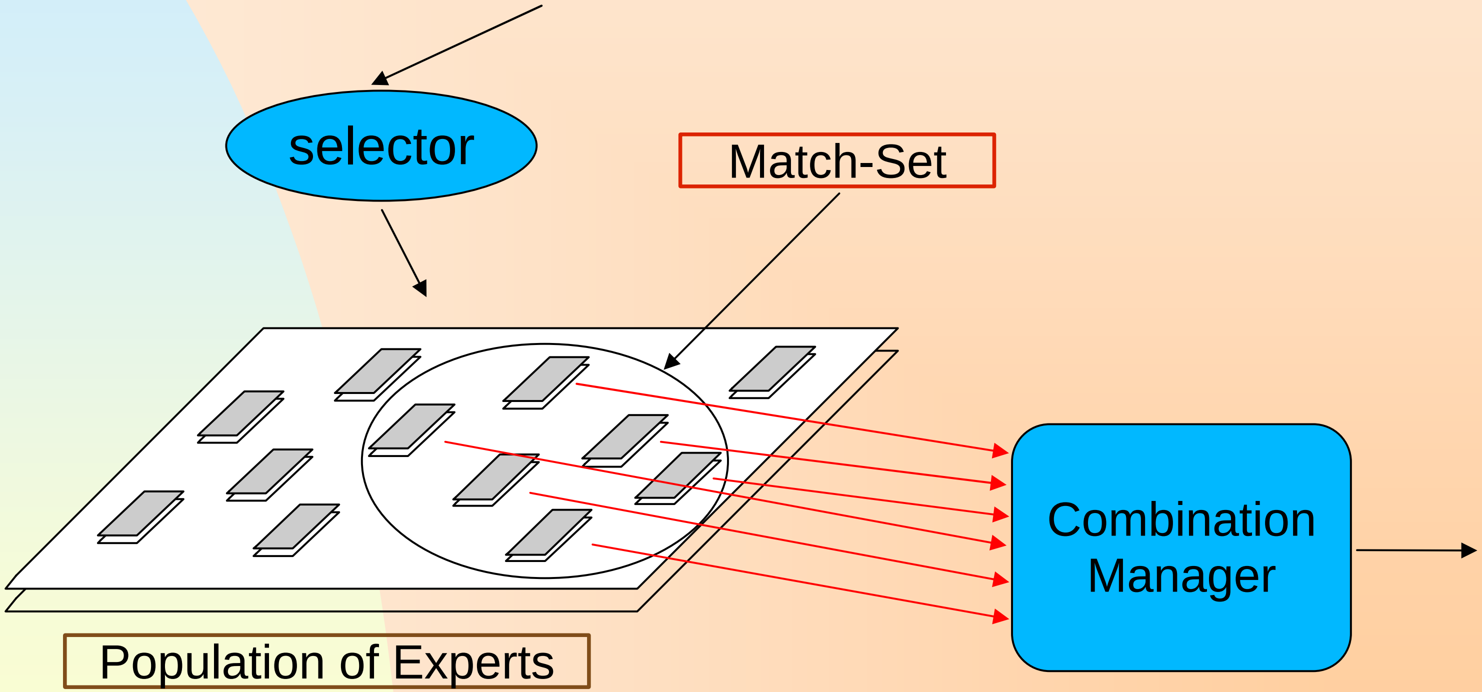


MASSP: Outputs Combination

- Given an input, several experts (in the match-set) concur to classify it
- Each expert outputs three real values in [0,1] for alpha-helices, beta-sheets, and coils
- Each expert concur in the voting activity depending on its strength

MASSP: Outputs Combination

`Input:` `... msgkmtgivkwfnad ...`



selector

Match-Set

Combination Manager

Population of Experts

MASSP: Selective Environment

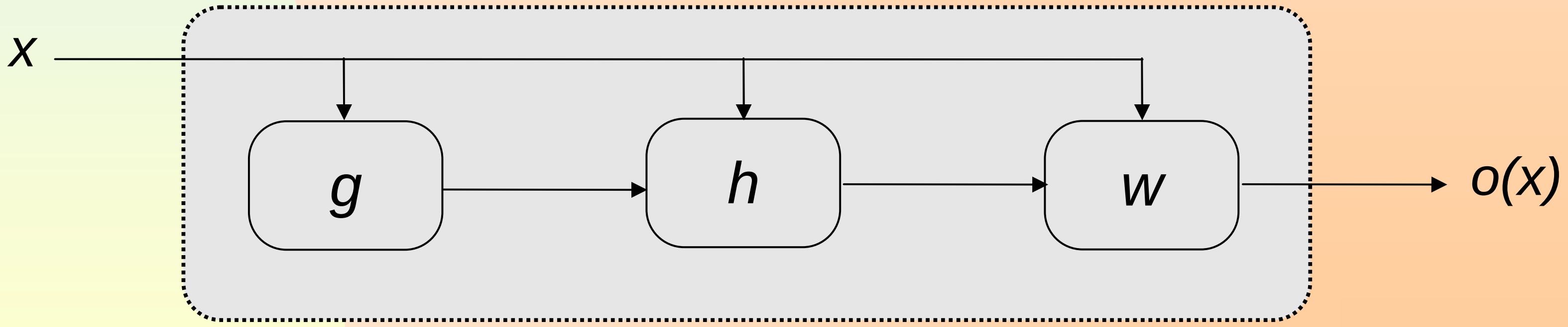
- The fitness of each expert is updated according to its performances over the training set, thereby enforcing dynamic adaptation to the given environment

MASSP Micro Architecture

The micro architecture is concerned with experts' “internals”

Each Expert Embodies ...

- A genetic classifier *g(_)* – the guard
- A feed-forward artificial neural network *h(_)* – the embedded classifier
- A weighting function *w(_)* – the modulator



For Each Classifier ...

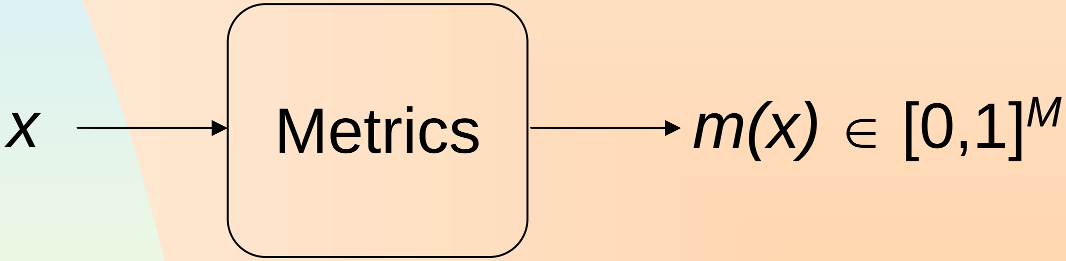
- The guard *g(_)* is devoted to control the activation of the embedded classifier *h(_) according to the (flexible) matching performed on the given input*
- The embedded classifier *h(_)* performs the actual classification
- The modulator *w(_)* is used to strengthen or weaken the embedded classifier's output according to the strength of the given expert

Guards

- Guards are entrusted with soft-partitioning the input space according to some domain knowledge, embodied in form of suitable metrics (*m*), whose combination is controlled by an embedded pattern (*e*) handled by the underlying “Darwinian” environment

Guards: Metrics

- For the sake of simplicity, metrics can be seen as a pre-processing activity performed on the given input *x*:



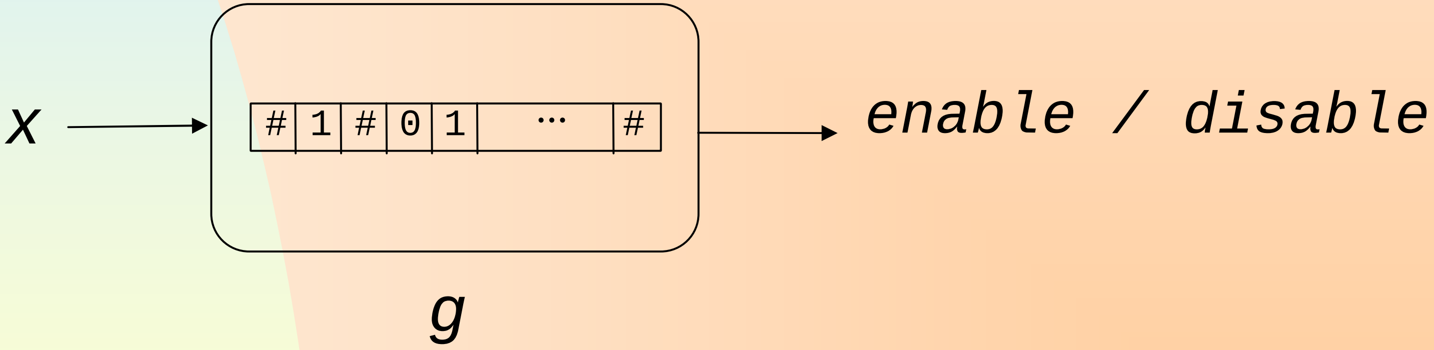
- A metric can be arbitrarily defined, provided that it is deemed biologically relevant
- A metric returns a result in [0,1]

Guards: “Biologically-Biased” Metrics

	Metrics	Rationale
1	Check whether hydrophobic amino acids occur in the window r according to a clear periodicity (e.g., one every 3-4 residues)	Sometimes hydrophobic amino acids are regularly distributed along alpha-helices
2	Check whether the window r contains numerous residues in {A,E,L,M} and few residues in {P,G,Y,S}	Alpha helices are often evidenced by {A,E,L,M} residues, whereas {P,G,Y,S} residues account for their absence
3	Check whether, on the average, the window r is positively charged or not	A positive charge might account for alpha helices or beta sheets.
4	Check whether, on the average, the window r is negatively charged or not	A negative charge might account for alpha helices or beta sheets
5	Check whether, on the average, the window r is neutral	A neutral charge might account for coils
6	Check whether the window r mostly contains “small” residues	Small residues might account for alpha helices or beta sheets
7	Check whether the window r mostly contains polar residues	Polar residues might account for alpha helices or beta sheets

Guards: Embedded Patterns

- An embedded pattern *e* is a string in $\{0,1,\#\}^M$
- *Embedded patterns are created and retained according to a Darwinian policy*



Guards: Flexible Matching

- *Let us denote with g(x) the result of flexible matching performed by an expert on the input x:*

  $g(x) = 1 – d(e,m(x))$

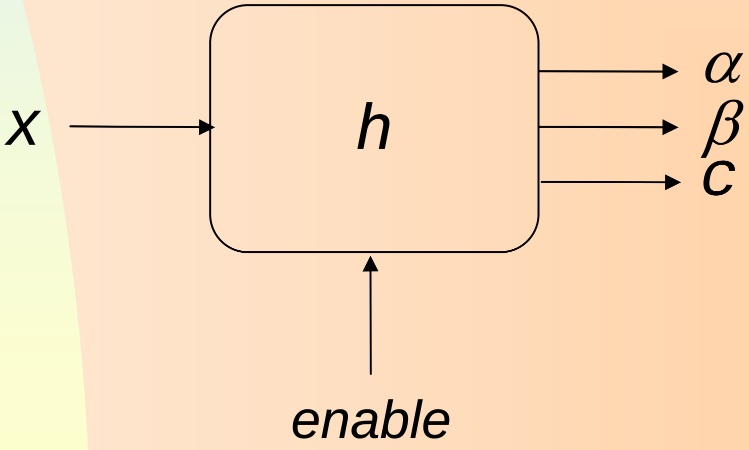
  $d(e,m(x)) = max_i (| e_i - m_i(x) |)$

  where i ranges over all metrics that are <u>not disregarded</u> by the embedded pattern

The adopted distance measure *d*(_,_) is the Minkowski's L_∞ metrics

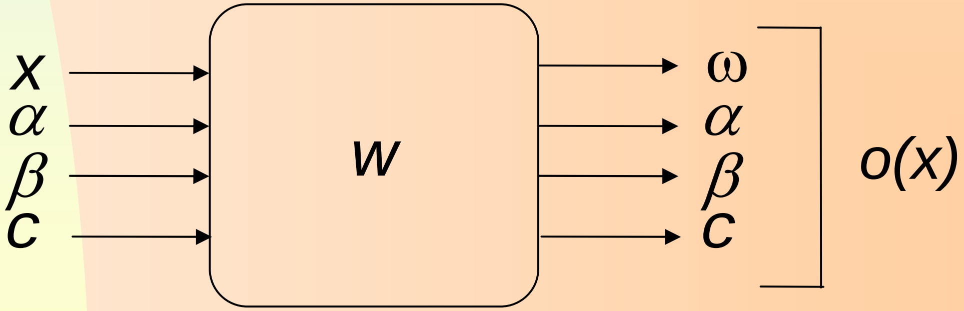
Embedded Classifiers

- An embedded classifier is entrusted with performing the actual classification task
- Each embedded classifier outputs three signals in [0,1] – one for each class label (i.e., alpha-helix, beta-sheet, and coil)



Modulator

- Evaluates the ability (ω) of the current expert to deal with the given input according to:
  - The expert fitness (handled by the genetic environment)
  - The result of flexible matching (handled by the guard)
  - The reliability of the prediction (that can be evaluated starting from α, β, c)



Experimental Results

Focus: impact of domain-specific metrics on the performances of the overall system

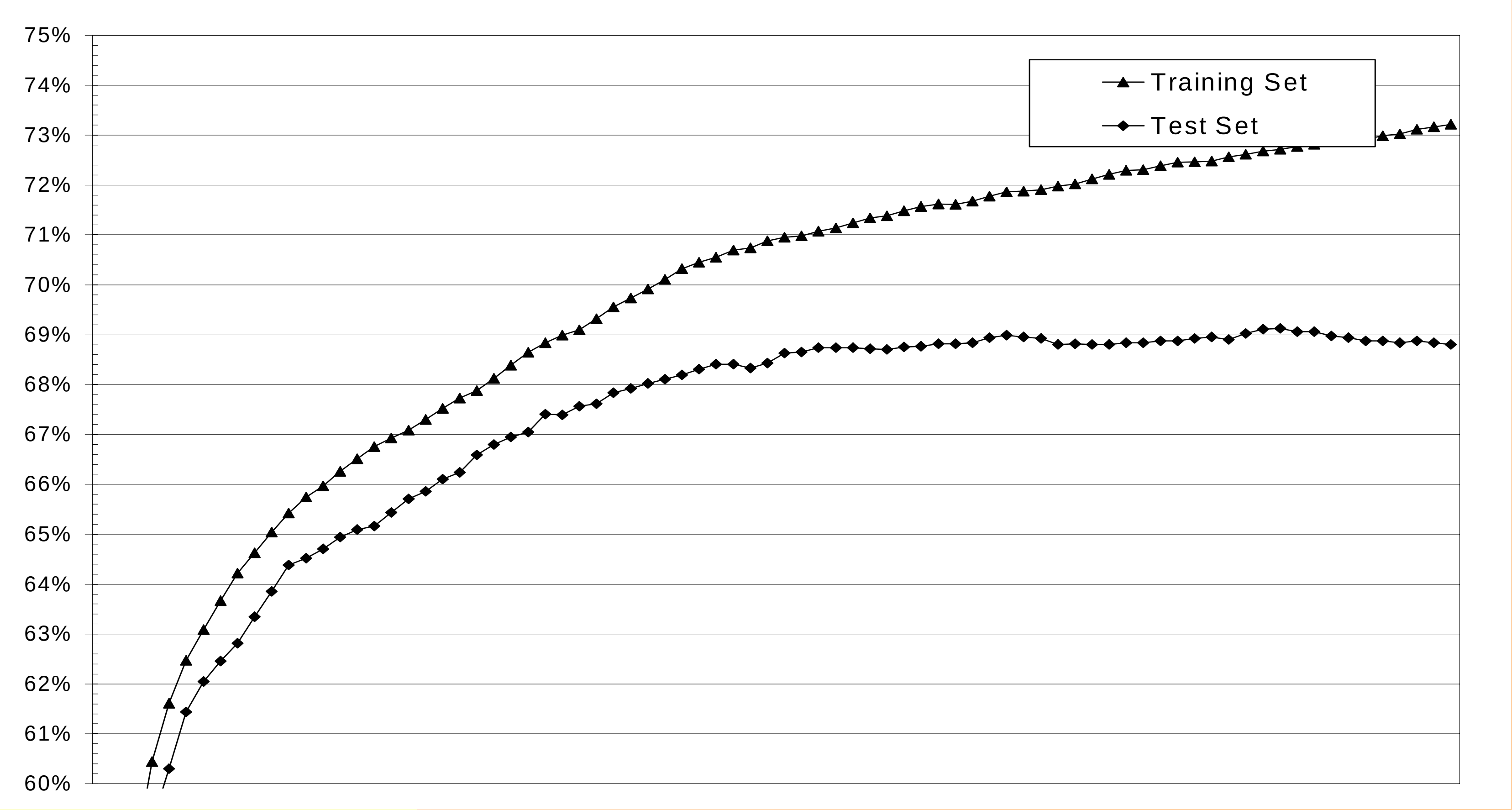
Experimental Results

- Experiments have been performed on the RS126 and CB396 datasets of proteins
- Focus: assessing the impact of the domain knowledge on the performance of the overall system.

Experimental Results

- Randomly-generated guards
  - 600 experts
  - ~20 experts (average) involved in the match set
  - 69.1% system precision on test set

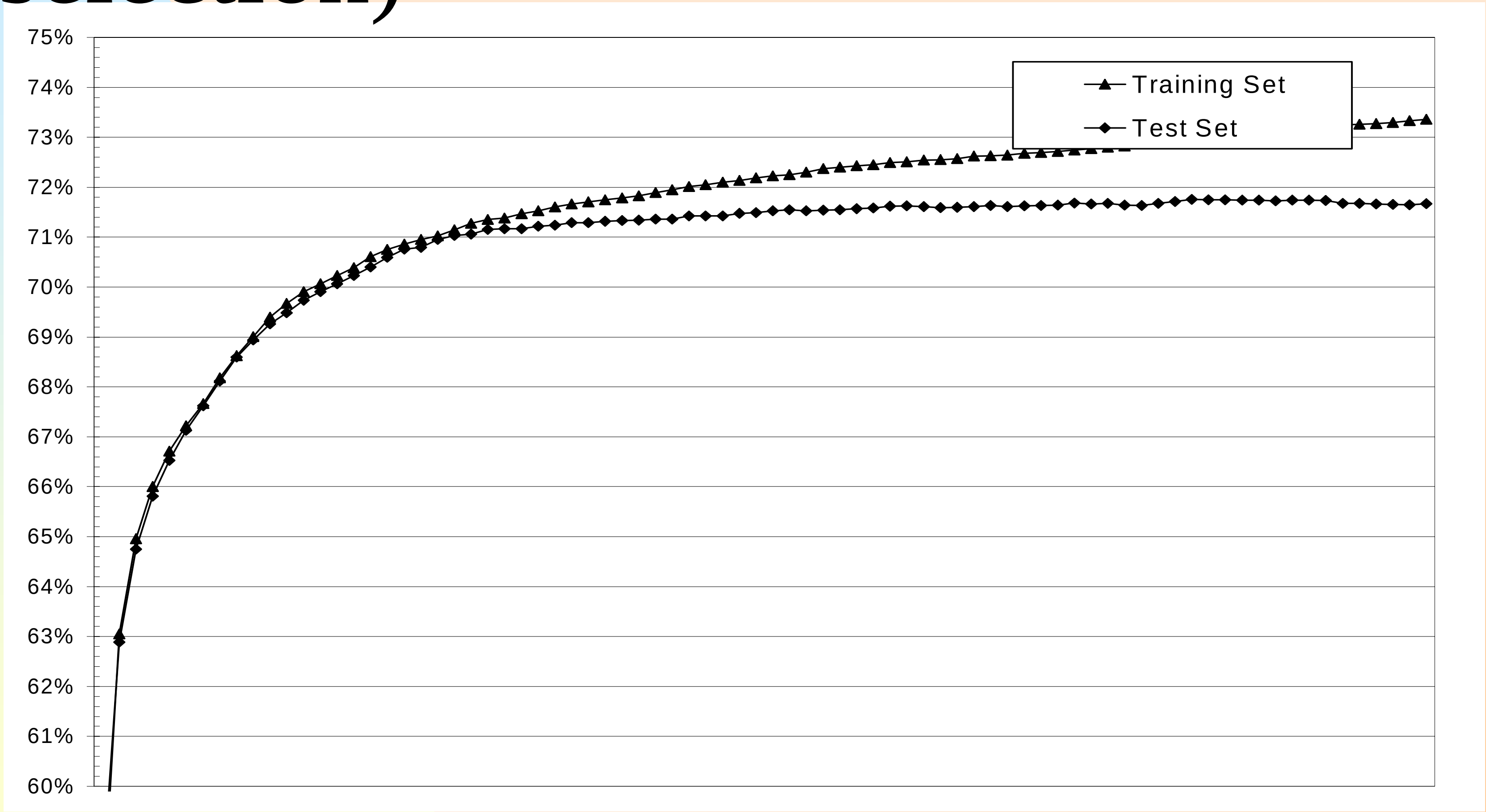
Experimental Results (with random generation)



Experimental Results

- guards obtained by enforcing a genetic selection
  - 600 experts
  - ~20 experts (average) involved in the match set
  - 71.8% system precision on test set

Experimental Results (with genetic selection)



Experimental Results

- improvement of ~2% with respect to the random case
  - In our opinion, the guards of the most successful experts embed combination of metrics that are effective in simplifying their learning task

Experimental Results

- MASSP vs other programs (7-fold cross validation)

	RS126	**CB396**
Method	**Q3**	**Q3**
PHD	73.5	71.9
DSC	71.1	68.4
PREDATOR	70.3	68.6
NNSSP	72.7	71.4
CONSENSUS	74.8	72.9
MASSP	71.7	69.5

Further Experimental Results

- After publishing the paper, we performed further tests using a more recent release of the system, characterized by:
  - Hybrid input encoding (Blosum80+multialignment)
  - Improved implementation of the adopted metrics
  - Two-tiered training for single experts, consisting of 5 epochs with global visibility + 10-20 epochs with local visibility (according to the guard)

Further Experimental Results

- We run the improved system on a larger dataset, used to train SSPRO (courtesy of G. Pollastri)
- The overall accuracy of the system is now 74.5

Conclusions and Future Work

Conclusions and Future work

- Multiple experts biased with relevant domain knowledge allow to improve the result over single (or "unbiased" multiple) experts
- Further improvements are expected depending on the adoption of:
  - More "biologically-biased" metrics
  - Metrics based on Hidden Markov Models
  - Recurrent ANN architectures (for embedded experts)

Thank You Very Much for Your Attention !