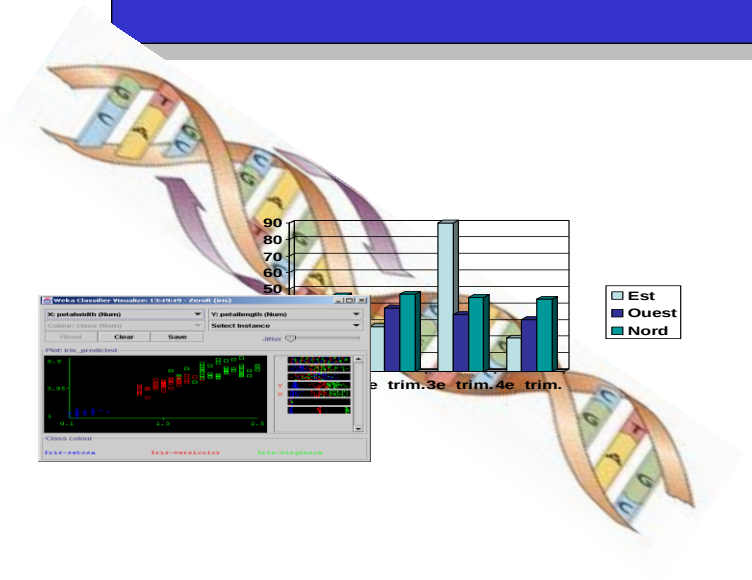
Protein sequences classification by means of feature extraction with substitution matrices



Rabie SAIDI
Mondher Maddouri
Engelbert Mephu Nguifo

JOBIM'10, Montpellier

Content

Families of sequences

F1	F2		Fn
Seq**1**.1	Seq**2**.1		Seq**n**.1
Seq**1**.2	Seq**2**.2		Seq**n**.2
…	…	…	…
Seq**1**.p	Seq**2**.p		Seq**n**.p

New sequence *S*

Biologist favourite tools

- Alignment
  - BLAST, FAST
- HMM profiles
  - HMMER, SAM
- A query belongs to the subject class with the best hit score
- Fundamentally depend on homology
  - Fail when classes contain distant sequences
- Non-discriminative (a score accompanied by e-value)

How to involve machine learning

Motif extraction selection

RAF
SSH
FFIT
MKGD
FYCG
VLAA
LVLQH
VMVWI
SYNTV

>SEQ 1
MEIPAVTEPSYNTVAKNDFMSGFLCFSINVRAFGITVPTPLYSLVFIIGVIGHVLVVLVLQHKRLRNMTSIYLFNLAISDLVFLSTLPFWV
DYIMKGDWIFGNAMCKFVSGFYYLGLYSDMFFITLLTIDRYLAVVHVVFALRARTVTFGIISSIITWVLAALVSIPCLYVFKSQMEFTYH
TCRAILPRKSLIRFLR

>SEQ 2
MAATASPQPLATEDADSENSSFYYYDRAFYLDEVAFMLCRKDAVVSFGKVFLPVFYSLIFVLGLSGNLLLLMVLLRYVPRRRMVEIYLL
NLAISNLLFLVTLPFWGISVAW MKGD HWVFGSFLCKMHCCFSPILYAFSSHRFRQYLKAFLAAVLGWHLA

>SEQ 3
MPTVASPLPLTTVSYNTVGSENSSSIYDYDYLDDMTILVCRKDEVLSFGRVFLPVVYSLIFVLGLAGNLLLLVVLLHSAPRRRTMELYLLN
LAVSNLLFVVTMPFWAISVAWHWVFGSFLCKVISTLYSINFYCGIFFITCMSLDKYLEIVHAQPLFYCGHRPKAQFRNLLLIVMVWITSL
AISVPEMLTLFLHSLLDLHVF MKGD GNCEISHRLDYTLVLQHLQVTESLAFSHCCFT

>SEQ 4
MPTIASPLPLATTGPENGSSIYDYDYLDDVTVLVCSRAFEVLSFGRVFLPVVYSLIFVLVLAAGLAGNLLLLVVLLHSVPQRRRMIELYLL
NLAVSNLLFVVTMPFWAISVAWHWVFGSFLCKVVSTLYSINFYCGIFFITCMSLDKYLEIVHAQPVMVWILHRPKTRFRNLLVWIT

Encoding

	RAF	SSH	FFIT	MKGD	FYCG	VLAA	LVLQH	VMVWI	SYNTV
SEQ 1	1	0	0	1	0	1	1	0	1
SEQ 2	1	1	0	1	0	0	0	0	0
SEQ 3	0	0	1	1	1	0	1	1	1
SEQ 4	1	0	0	0	1	1	0	1	0

- Data: protein sequences
  - Alphabet: 20 amino acids
  - Format inadequate for ML ou DM
  - → Preprocessing is needed
- Preprocessing:
  - Looking for descriptors / motifs
  - Using them as attributes
  - structural and functional importance of preserved regions
  - These regions can be used as descriptors for the bio-sequences [Nevill-Manning et al, 1998]
- Loss of information ?
  - Due to format change
- Efficient preprocessing → better results
  - Exp: accuracy in classification

- N-Grams [Shannon, 1951]
- Active Motifs [Wang et al, 1999]
- Discriminative Descriptors [Maddouri et al, 2004]
- Amino Acid Composition [Zhang et al, 2003]
- Functional Domain Composition [Yu et al, 2006]

Motivation

- The mentioned methods neglect the substitution

- Some amino acids have similar proprieties
  - Some substitution can be without effect on the function nor the structure of the protein

  → Same thing can be deduced for features

- Idea : use substitution matrices to define similarity between features
  - Only one of the substitutable features is kept
  - Number of features will be reduced
  - Any impact on classification?

- Let M and M' 2 motifs composed of k amino acids
- For each amino acid of M with index i, its probability P_i to not mutate is :

$$P_i = S(M[i], M[i]) / \sum_{j=1}^{20} S^{+}(M[i], AA_j)$$

  - AA_j is the amino acid with index j among the 20 amino acids
- The probability P_m that M mutate to other motifs, is :

$$P_m(M) = 1 - \prod_{i=1}^{k} P_i$$

  - $P_m(\mathrm{LLK}) = 1 - (4/9 * 4/9 * 5/9) = 0.89$

- We note by S_m (M, M') the score of substitution of a motif M' by a motif M :

$$S_m\,(M, M') = \sum_{i=1}^{k} S(M[i], M'[i])$$

  - S_m(LLK, VMK) = 8
  - S_m(LLK, LLK) = 13

- We note by PS (M, M') the probability of substitution of a motif M' by a motif M

$$PS\,(M, M') = S_m\,(M, M') \,/\, S_m\,(M, M)$$

  - PS(LLK, VMK) = 8 / 13 = 0.61

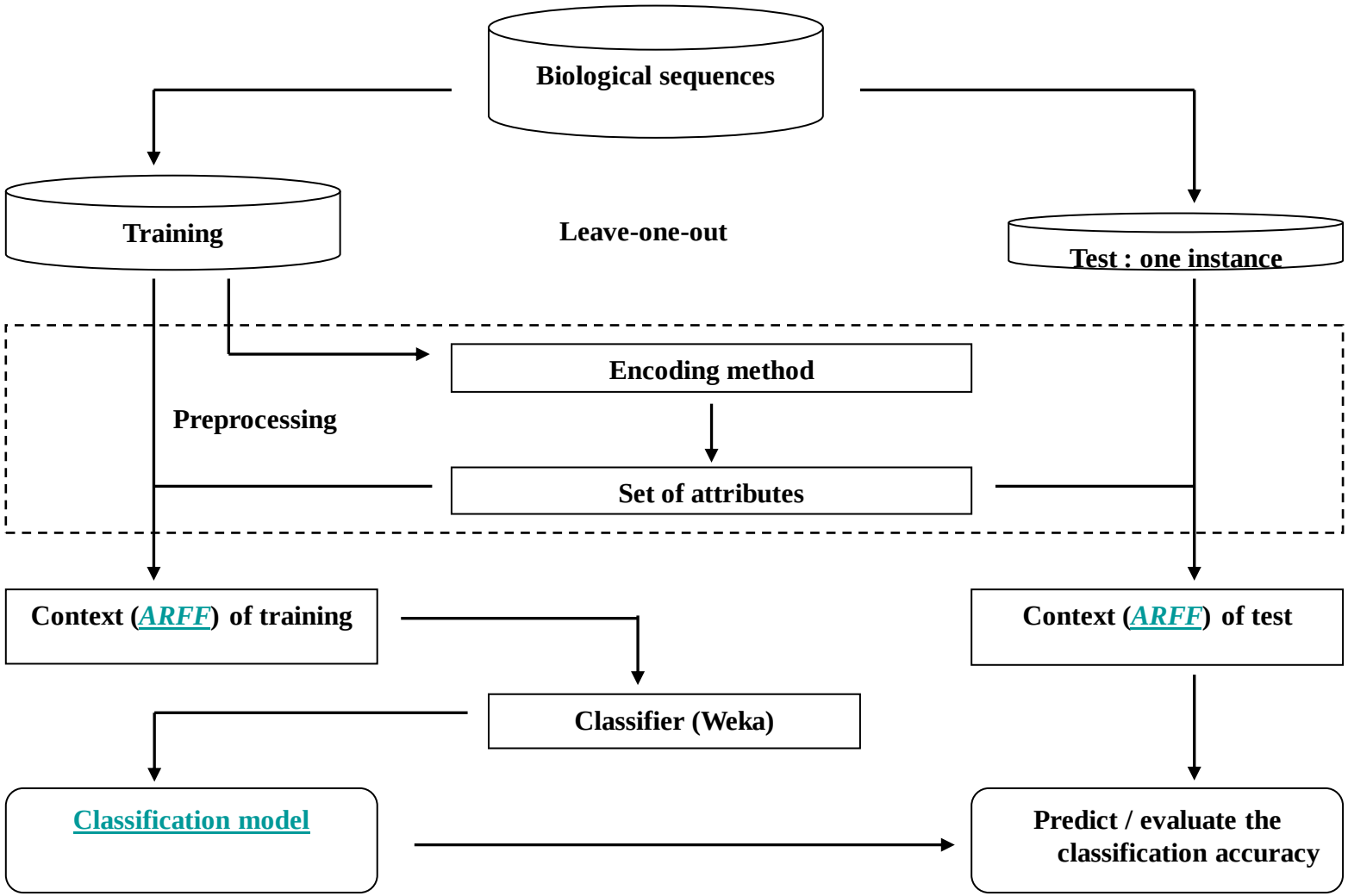
- A motif *M* substitutes a motif *M'* if :
  - *M* and *M'* have the same length *k*,
  - $S(M[i], M'[i]) >= 0$, $i = 1..k$,
  - $PS(M, M') >= T$, where *T* is a user-specified threshold : $0 \leq T \leq 1$.

- Research of repeated words (adaptation of the KMR algorithm [Karp et al, 1972])
  - Equivalence notion
- Discrimination : ***x*** is discriminative of a family ***fi***
  - Occurrence rate of ***x*** in ***fi*** **>= *α***
  - Occurrence rate de ***x*** in others ***f*** **<= *β***
- Minimality
  - A sub-sequence is called minimal if it does not contain other discriminative sub-sequences.
- Example

- Identification of main motifs and filtering
  - The motifs are clustered in groups
  - Group : a ***main motif*** *MM* + other motifs
  - *MM* substitutes all the motifs in its group
  - *MM*: the most likely motif in its cluster to mutate to other motifs i.e having the highest *Pm* in its group
  - Filtering : keep only *MM*
  - Example

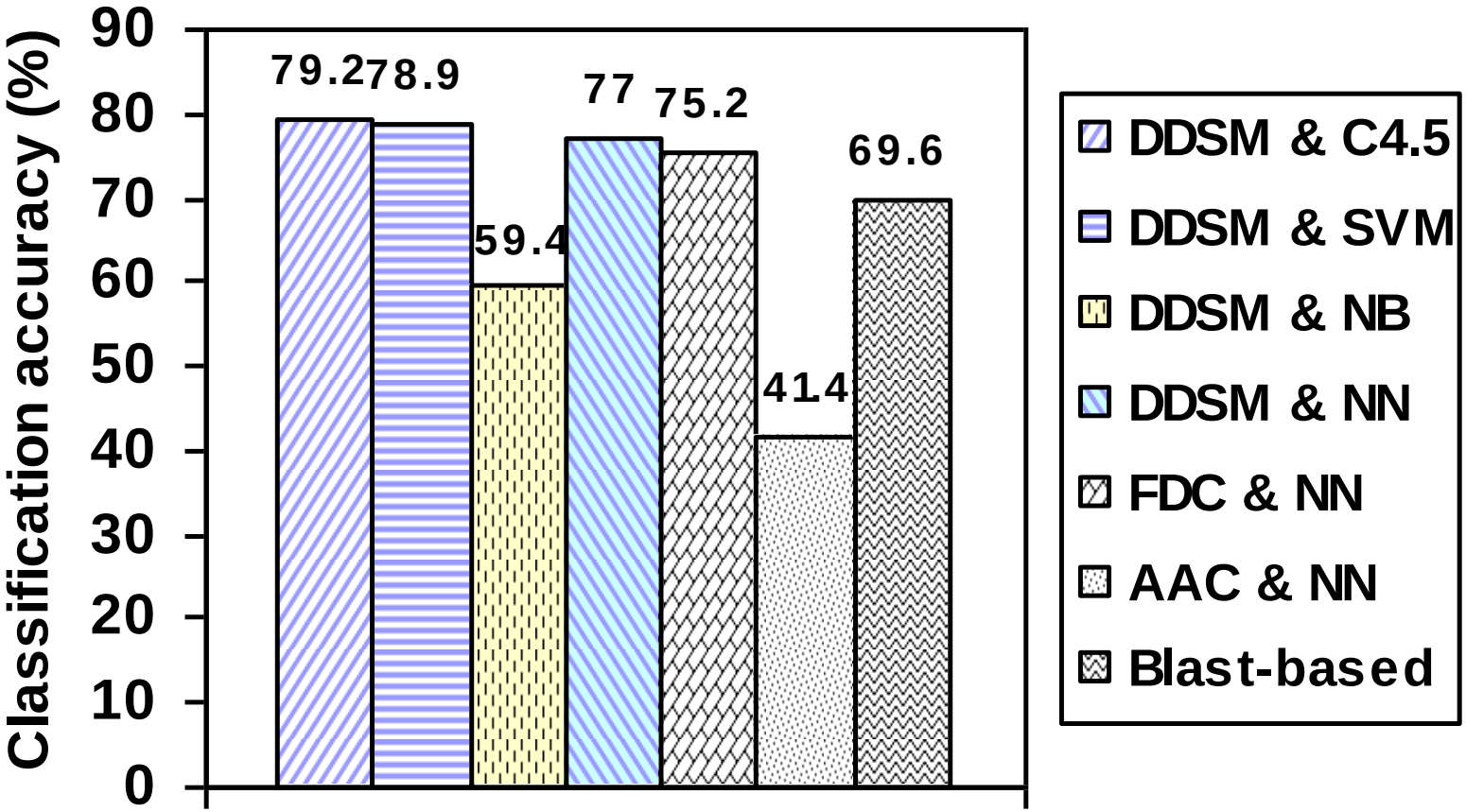
- Construction of the context (binary vectors)
  - Compare each motif with the k-grams (k=length of motif) of each sequence
    - Until find a substitute and note 1 in the binary table
    - Or cross the whole sequence without finding a substitute and note 0
  - → Richer context

Protocol

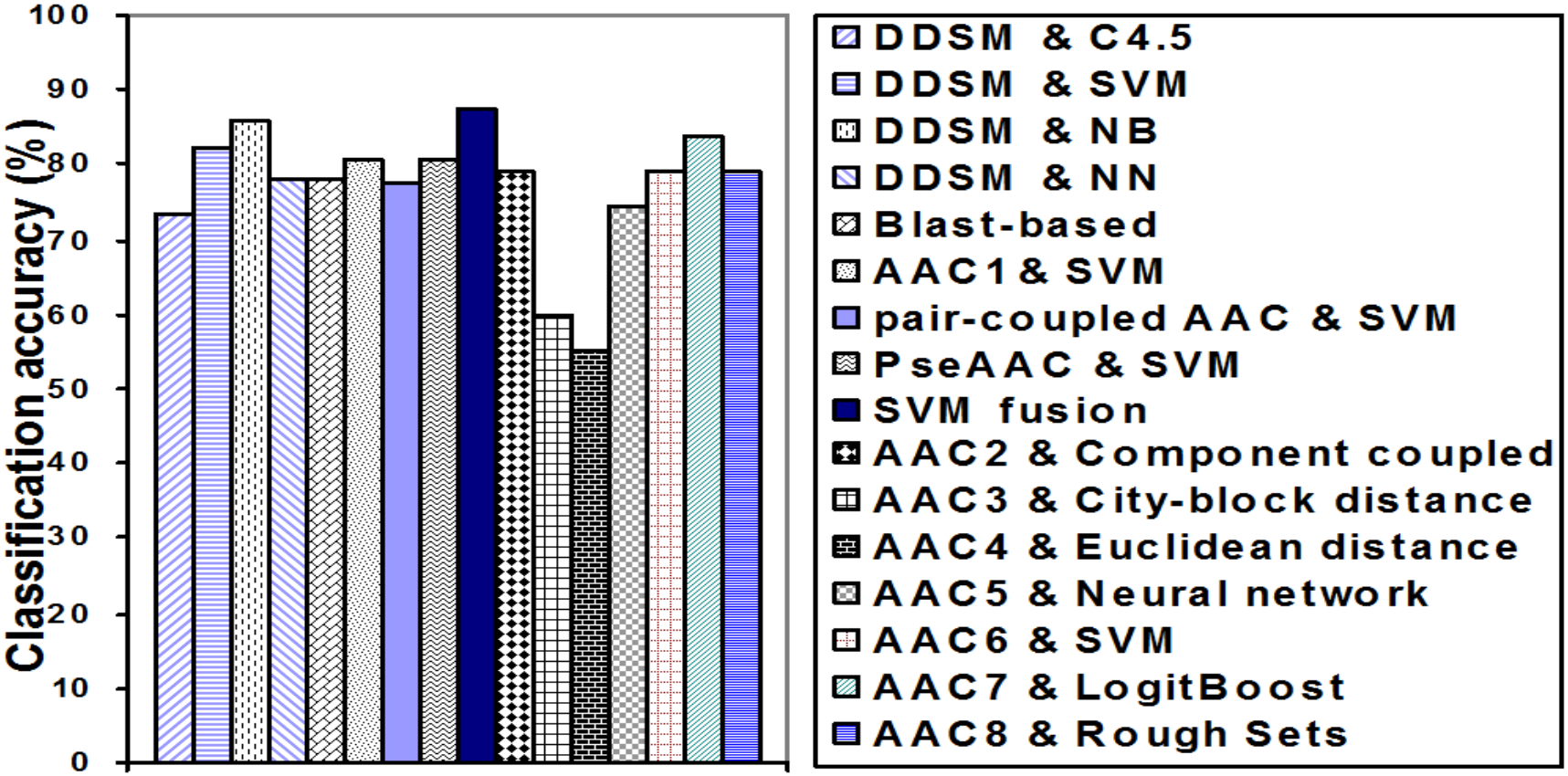

Biological sequences
Training
Leave-one-out
Test : one instance
Encoding method
Preprocessing
Set of attributes
Context (ARFF) of training
Context (ARFF) of test
Classifier (Weka)
Classification model
Predict / evaluate the classification accuracy

Dataset (source)	Identity percentage	Family/class	Size	Total
DS1 (Yu et al, 2006) *BMC Bioinformatics*	25 %	Monomer	208	717
		Homodimer	335	
		Homotrimer	40	
		Homotetramer	95	
		Homopentamer	11	
		Homohexamer	23	
		Homooctamer	5	
DS2 (Zhou et al, 2006) *Anal Biochim*	84 %	All-α domain	70	277
		All-β domain	61	
		α / β domain	81	
		α + β domain	65	

Classification results – DS1



Classification results – DS2



Substitution matrix effect – DS1 & DS2

Substitution matrix	Attributes	Accuracy (%)			
		C4.5	SVM	NB	NN
Blosum45	377	78.5	79.2	59.4	77.7
Blosum62	508	79.2	78.9	59.4	77
Blosum80	532	77.6	80.5	60	77.6
Pam30	2873	77.8	82	60.3	76.7
Pam70	802	78.1	80.5	60.5	77
Pam250	1123	77.3	79.4	59.6	78.7

Substitution matrix	Attributes	Accuracy (%)			
		C4.5	SVM	NB	NN
Blosum45	2603	69.3	82.3	85.9	78
Blosum62	3083	73.3	82.3	85.9	78
Blosum80	3146	70.1	82.3	84.1	78
Pam30	3830	69.3	82.3	84.5	78
Pam70	3822	70.4	82.3	84.5	78
Pam250	969	66.1	85.2	79.4	78

Conclusion

- Motif-based encoding
  - May allow reliable description of protein
  - Allow the injection of external information (pH, temperature,…)

- DDSM (discriminative descriptors with substitution matrix)
  - A discriminative enconding method taking the substitution into account
  - Low number of features
  - Helps with classification task
  - Coupled with SVM : effcient protein classifier

- Blosum matrices with higher numbers and Pam matrices with low numbers allow the building of fewer features

- Variances of accuracies are slight when varying the substitution matrices with the same classifier

- More can be found in [Saidi et al., BMC Bioinformatics 2010]

Research article

Open Access

Protein sequences classification by means of feature extraction with substitution matrices

Rabie Saidi[1,2,3,4], **Mondher Maddouri**[4,5] and **Engelbert Mephu Nguifo**[1,2]

1 LIMOS - Blaise Pascal University - Clermont University, BP 10448, Clermont-Ferrand 63000, France
2 LIMOS - CNRS UMR 6158, Aubière 63173, France
3 Department of Computer Science - FSJ - University of Jendouba, UMA Street, Jendouba 8100, Tunisia
4 URPAH - FST - University of Tunis El Manar, Academic Campus, Tunis 2092, Tunisia
5 Department of Computer Science - FSG - University of Gafsa, Campus of Sidi Ahmed Zarroug, Gafsa 2112, Tunisia

author email corresponding author email

BMC Bioinformatics 2010, **11**:175 doi:10.1186/1471-2105-11-175

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2105/11/175

Received: 4 September 2009
Accepted: 8 April 2010
Published: 8 April 2010

BMC Bioinformatics
Volume 11

Viewing options:
- Abstract
- **Full text**
- PDF (947KB)
- Additional files

Associated material:
- Readers' comments
- PubMed record

Related literature:
- Articles citing this article
  on Google Scholar
  on PubMed Central
- Other articles by authors
  on Google Scholar
  on PubMed
- Related articles/pages
  on Google
  on Google Scholar
  on PubMed

THANKS

This work was partially supported by the French-Tunisian project CMCU-Utique 05G1412 and the LifeGrid PREFON_META project