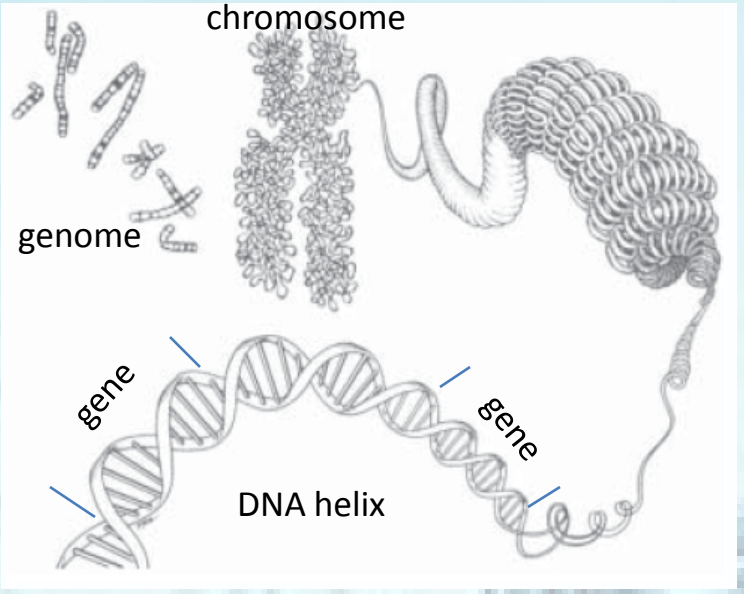
CONSTRUCTION METRICS FOR BIOMOLECULAR SEQUENCES

Valentina V. Sulimova [1] vsulimova@yandex.ru
Oleg S. Seredin [1] oseredin@yandex.ru
Vadim V. Mottl [2] vmottl@yandex.ru

[1] *Tula State University*
[2] *Moscow Computing Centre of the RAS*

IDP-2016, Barcelona, October 2016

TYPES OF BIOMOLECULAR SEQUENCES

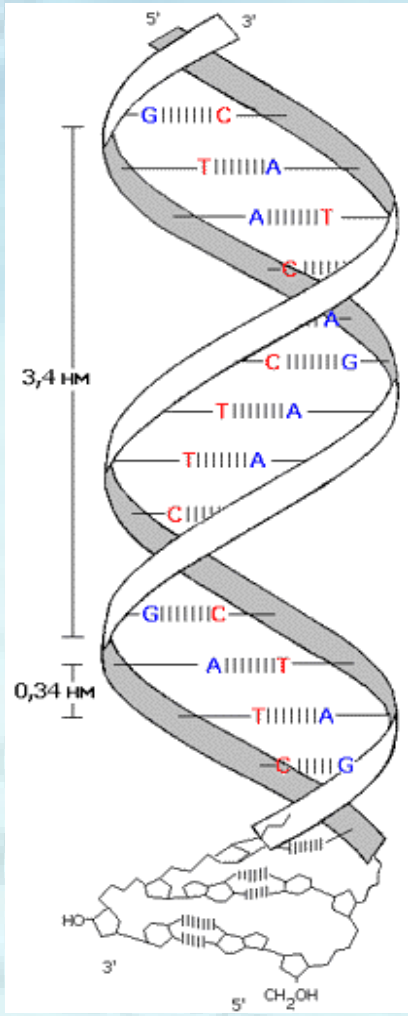


Nucleotid sequences (DNA) – symbolic sequences over 4 nucleotid s

a	adenine	c	cytosine
g	guanine	t	thymine

An example of a fragment of a nucleotid sequence

gatttggggttcaaagcagttatcgatcaaataatcc
atttgtetcaactcacgtttcaaagcatcgatcaaag
atttggggttcaaagcagttatcgatcaaataatcca
tttgtetcaat



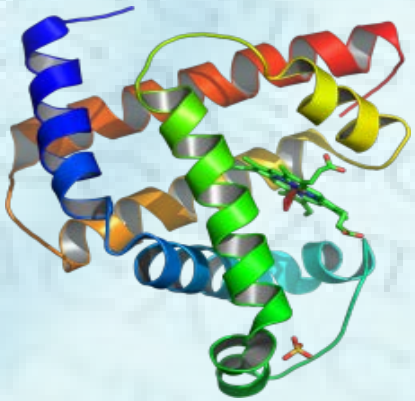
DNA double helix, which defines amino acid sequences

Amino acid sequences (proteins) – symbolic sequences over 20 amino acids

Alanine	Ala	A	Methionine	Met	M
Cysteine	Cys	C	Asparagine	Asp	N
Aspartic	Asp	D	Proline	Pro	P
Glutamic	Glu	E	Glutamine	Gln	Q
Phenylalanine	Phe	F	Arginine	Arg	R
Glycine	Gly	G	Serine	Ser	S
Histidine	His	H	Threonine	Thr	T
Isoleucine	Iso	I	Valine	Val	V
Lysine	Lys	K	Tryptophan	Trp	W
Leucine	Leu	L	Tyrosine	Tyr	Y

An example of an amino acid sequence

MLDEQLAWAYACLKHGRELPTDDILMSTSEKLSQQLVIKLIEVIKCIEKDGIFSRILK
GVADAVCLKAQFLRGMITLKRTPCSLPMYTLFVYVLTIPTLRTRVIRDPLLTQCKDV
VLKYQPGDCITLLKAALNCHQCNKDCDKCKYILDPLLGQTHRTKGVFFVCEEQLA
WAYACLKHGRELPTDDILMSTSEKLSQQLVIKLIEVIKCIERILKGVADAMYTLFVYV
LTIPTLRTRVIRDPLLTQCKDVVLKYQPGDCITLLKAALNCHQ



An example of a protein's space structure

COMPARING BIOMOLECULAR SEQUENCES

For successful biological sequences analysis the comparing measure should :

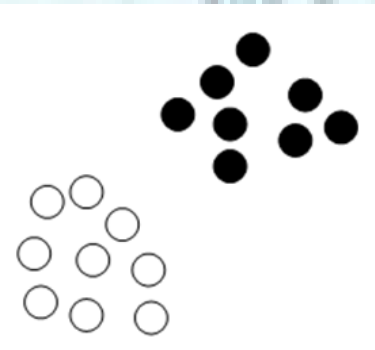
1) form a space, satisfying the compactness hypothesis
2) have low computational complexity
3) allow for applying effective and convenient SVM-based methods

---

Typical example of biological task:
Viewing amino acid sequence, to determine, does the respective protein is regulator of destructor

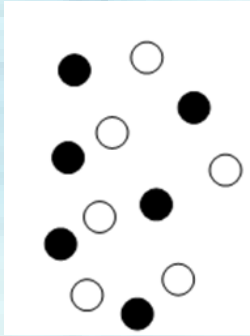
---

Space is adequate for biological task (the compactness hypothesis holds true)



Proteins, performing the same function, are mapped into compact sets of points

Space is NOT adequate for biological task (the compactness hypothesis holds NOT true)



Proteins of different types are «mixed»

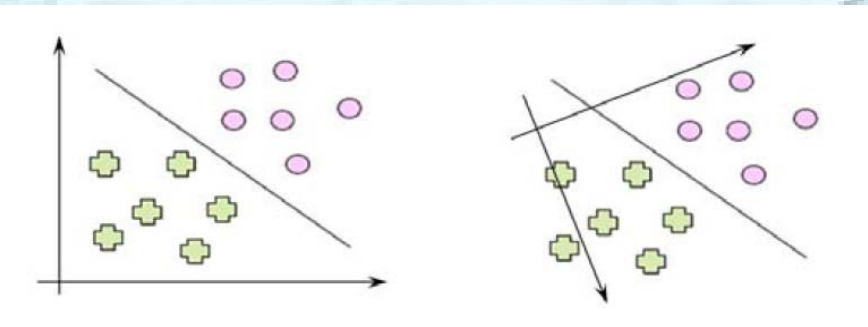
- regulators
- destructors

PROPERTIES OF COMPARING MEASURES

Types of comparing measures	Comp. complexity	Adequacy for biological tasks	Suitability for SVM-based methods
Alignment-based similarity measures: Needleman-Wunsch (NW) alignment, Smith-Waterman (SW) alignment etc.	medium	+	-
Secondary features on the basis of alignment-based similarity measures	medium	+ -	+ but with missing computational advantages of SVM
Evolutionary-based kernels	high and extra high	+	+
Another kernels String, diffusion, FFT, Spectrum, etc.	different: from low to high	-	+

KERNEL PROPERTIES ARE EXECCIVE

1) A metric (i.e. relative positions of objects), but not object's coordinates define the result of analysis
2) There are classes of kernels, defining the same metric and so the same decision rules
3) There are metric-based versions of SVM*



Orientation and position of an optimal hyperplane , separating objects of two classes, **depend only on a metric** and don't depend on a centre and an orthonormal basis of a linear space

Metric - function $\rho(x, y)$, such as:

$$1.\ \rho(x, y) \geq 0$$

$$2.\ \rho(x, y) = 0 \Leftrightarrow x = y$$

$$3.\ \rho(x, y) = \rho(y, x)$$

$$4.\ \rho(x, y) + \rho(y, z) \geq \rho(x, z)$$

**Abramov V.I., Seredin O.S., Mottl V.V. Pattern recognition training with support objects method in Euclidean metric spaces with affine operations// Transactions TSU. Natural Sciences, Tula, 2013, V. 2, Part 1, pp. 119-136. (In Russian)*
**Seredin O.S. Mottl V.V. Method of support objects for training in metric spaces of arbitrary kind // Transactions of TSU. Natural Sciences. Tula, 2015, V. 4. pp.49-66 (in Russian)*

PROPERTIES OF COMPARING MEASURES

Types of comparing measures	Comp. complexity	Adequacy for biological tasks	Suitability for SVM-based methods
Alignment-based similarity measures: Needleman-Wunsch (NW) alignment, Smith-Waterman (SW) alignment etc.	medium	+	-
Secondary features on the basis of alignment-based similarity measures	medium	+ -	+ but with missing computational advantages of SVM
Evolutionary-based kernels	high and extra high	+	+
Another kernels String, diffusion, FFT, Spectrum, etc.	different: from low to high	-	+
Algebraic metrics	low, medium	-	+

PROPERTIES OF COMPARING MEASURES

Types of comparing measures	Comp. complexity	Adequacy for biological tasks	Suitability for SVM-based methods
Alignment-based similarity measures: Needleman-Wunsch (NW) alignment, Smith-Waterman (SW) alignment etc.	medium	+	-
Secondary features on the basis of alignment-based similarity measures	medium	+ -	+ but with missing computational advantages of SVM
Evolutionary-based kernels	high and extra high	+	+
Another kernels String, diffusion, FFT, Spectrum, etc.	different: from low to high	-	+
Algebraic metrics	low, medium	-	+
The proposed alignment-based metric	medium	+	+

METRICS ON THE SET OF AMINO ACIDS

$A = \{\alpha^1, ..., \alpha^m\}$, $m = 20$ - set of amino acids

Theoretical conception of amino acid's comparison :

Probabilistic model of evolution of amino acids PAM (Point Accepted Mutation) by M. Dayhoff

The main notion:

Markov chain of evolution of amino acids in some point of a protein sequence with matrix of transitional probabilities $\Psi_{[1]} = \left(\psi_{[1]}\left(\alpha^j \mid \alpha^i\right)\right)$

Suppositions:

$\xi(\alpha^j) = \sum_{\alpha^i \in A} \xi(\alpha^i)\psi_{[1]}\left(\alpha^j \mid \alpha^i\right)$ - ergodicity with final distribution $\xi(\alpha^i), i = 1, ..., m$

$\xi(\alpha^i)\psi_{[1]}\left(\alpha^j \mid \alpha^i\right) = \xi(\alpha^j)\psi_{[1]}\left(\alpha^i \mid \alpha^j\right)$ - reversibility

METRICS ON THE SET OF AMINO ACIDS

$A=\{\alpha^1,...,\alpha^m\},\ m=20$ - set of amino acids

Theoretical conception of amino acid's comparison :

Probabilistic model of evolution of amino acids PAM (Point Accepted Mutation) by M. Dayhoff

The main notion:

Markov chain of evolution of amino acids in some point of a protein sequence with matrix of transitional probabilities $\Psi_{[1]}=\left(\psi_{[1]}\left(\alpha^j\,|\,\alpha^i\right)\right)$

Suppositions:

$\xi(\alpha^j)=\sum_{\alpha^i\in A}\xi(\alpha^i)\psi_{[1]}\left(\alpha^j\,|\,\alpha^i\right)$ - ergodicity with final distribution $\xi(\alpha^i), i=1,...,m$

$\xi(\alpha^i)\psi_{[1]}\left(\alpha^j\,|\,\alpha^i\right)=\xi(\alpha^j)\psi_{[1]}\left(\alpha^i\,|\,\alpha^j\right)$ - reversibility

Theorem 1*.

For any $\Psi_{[s]}=\underbrace{\left[\Psi_{[1]}\times...\times\Psi_{[1]}\right]}_{s}$ similarity measure $\kappa_s(\alpha^i,\alpha^j)=\psi_{[s]}(\alpha^i\,|\,\alpha^j)\big/\xi(\alpha^i)$

is a **kernel function** (forms nonnegative matrix for amino acids $[\kappa_s(\alpha^i,\alpha^j), i,j=1,...,m]$)

**Sulimova V.V. Kernel functions for signals and symbolic sequences of different length. PhD thesis. 2009 (In Russian)*

METRICS ON THE SET OF AMINO ACIDS

$A=\{\alpha^1,...,\alpha^m\},\ m=20$ - set of amino acids

Theoretical conception of amino acid's comparison :

Probabilistic model of evolution of amino acids PAM (Point Accepted Mutation) by M. Dayhoff

The main notion:

Markov chain of evolution of amino acids in some point of a protein sequence with matrix of transitional probabilities $\Psi_{[1]}=\left(\psi_{[1]}\left(\alpha^j\,|\,\alpha^i\right)\right)$

Suppositions:

$\xi(\alpha^j)=\sum_{\alpha^i\in A}\xi(\alpha^i)\psi_{[1]}\left(\alpha^j\,|\,\alpha^i\right)$ - ergodicity with final distribution $\xi(\alpha^i), i=1,...,m$

$\xi(\alpha^i)\psi_{[1]}\left(\alpha^j\,|\,\alpha^i\right)=\xi(\alpha^j)\psi_{[1]}\left(\alpha^i\,|\,\alpha^j\right)$ - reversibility

Theorem 1*.

For any $\Psi_{[s]}=\underbrace{\left[\Psi_{[1]}\times...\times\Psi_{[1]}\right]}_{s}$ similarity measure $\kappa_s(\alpha^i,\alpha^j)=\psi_{[s]}(\alpha^i\,|\,\alpha^j)\big/\xi(\alpha^i)$

is a **kernel function** (forms nonnegative matrix for amino acids $[\kappa_s(\alpha^i,\alpha^j), i,j=1,...,m]$)

$\rho(\alpha^i,\alpha^j)=\left(\kappa_s(\alpha^i,\alpha^i)+\kappa_s(\alpha^j,\alpha^j)-2\kappa_s(\alpha^i,\alpha^j)\right)^{1/2}\ \forall s, s=1,2,...$ is **Euclidean metric****

**Sulimova V.V. Kernel functions for signals and symbolic sequences of different length. PhD thesis. 2009 (In Russian)*

***Mottl V.V. Metric spaces, enabling introducing linear operations and inner product // Reports of the RAS, 2003, V. 38. pp.1-4 (in Russian)*

ALIGNMENT OF SYMBOLIC SEQENCES

Ω - set of all sequences over alphabet $A = \{\alpha^1, \ldots, \alpha^m\}$

$\omega' = (\alpha'_1, \ldots, \alpha'_{N'}) \in \Omega$, $\omega'' = (\alpha''_1, \ldots, \alpha''_{N''}) \in \Omega$ - two sequences of different lengths N' and N''

Alignment is a way of arranging the sequences by inserting «gaps»

α'_1	α'_2	–	α'_3	α'_4	α'_5	α'_6	α'_7
	|		|	|			|
–	α''_1	α''_2	α''_3	α''_4	–	–	α''_5

An example of alignment

$\mathbf{w}$:

$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_4$	$\mathbf{w}_5$	$\mathbf{w}_6$	$\mathbf{w}_7$	$\mathbf{w}_8$
1	2	0	3	4	5	6	7
0	1	2	3	4	0	0	5

and it's mathematical representation as a table

Permissible alignment - alignment without two gaps in one position $\{i : \mathbf{w}_{i,1} = \mathbf{w}_{i,2} = 0\} = \varnothing$

$W_{N'N''}$ - set of permissible alignments of two sequences of lengths N' and N''

ALIGNMENT OF SYMBOLIC SEQENCES

Ω - set of all sequences over alphabet $A=\{\alpha^1,\dots,\alpha^m\}$

$\omega'=(\alpha'_1,\dots,\alpha'_{N'})\in\Omega,\quad \omega''=(\alpha''_1,\dots,\alpha''_{N''})\in\Omega$ - two sequences of different lengths N' and N''

Alignment is a way of arranging the sequences by inserting «gaps»

α'_1	α'_2	–	α'_3	α'_4	α'_5	α'_6	α'_7
	|		|	|			|
–	α''_1	α''_2	α''_3	α''_4	–	–	α''_5

An example of alignment

$\mathbf{w}:$

$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_4$	$\mathbf{w}_5$	$\mathbf{w}_6$	$\mathbf{w}_7$	$\mathbf{w}_8$
1	2	0	3	4	5	6	7
0	1	2	3	4	0	0	5

and it's mathematical representation as a table

Permissible alignment - alignment without two gaps in one position $\{i:\mathbf{w}_{i,1}=\mathbf{w}_{i,2}=0\}=\varnothing$

$W_{N'N''}$ - set of permissible alignments of two sequences of lengths N' and N''

Extended alphabet: $\tilde{A}=A\cup\{-\}=\{\alpha^1,\dots,\alpha^m,-\}=\{\tilde{\alpha}^1,\dots,\tilde{\alpha}^{m+1}\}$

Extended sequences: $\tilde{\omega}'=(\tilde{\alpha}'_1,\dots,\tilde{\alpha}'_{N_\mathbf{w}})\in\tilde{\Omega},\ \tilde{\omega}''=(\tilde{\alpha}''_1,\dots,\tilde{\alpha}''_{N_\mathbf{w}})\in\Omega$ of the same length $N_\mathbf{w}$

$$\tilde{\alpha}'_{\mathbf{w},i}=\begin{cases}\alpha'_{\mathbf{w}_{i,1}},\mathbf{w}_{i,1}\neq 0\\ -,\mathbf{w}_{i,1}=0\end{cases},i=1,\dots,N_\mathbf{w},\quad \tilde{\alpha}''_{\mathbf{w},i}=\begin{cases}\alpha''_{\mathbf{w}_{i,2}},\mathbf{w}_{i,2}\neq 0\\ -,\mathbf{w}_{i,2}=0\end{cases},i=1,\dots,N_\mathbf{w}.$$

METRIC OVER THE EXTENDED ALPHABET

$\rho(\alpha',\alpha'')$ - the metric over the initial alphabet $A=\{\alpha^1,\ldots,\alpha^m\}$

Extended alphabet: $\tilde{A}=A\cup\{-\}=\{\alpha^1,\ldots,\alpha^m,-\}=\{\tilde{\alpha}^1,\ldots,\tilde{\alpha}^{m+1}\}$

Extension of the metric:

$$\tilde{\rho}(\alpha',\alpha'')=\rho(\alpha',\alpha'')\ \forall\ \alpha',\alpha''\in A,$$
$$\tilde{\rho}(-,-)=0.$$

Theorem 2*.

Function $\tilde{\rho}(\alpha',\alpha'')$ is a metric if

$$\tilde{\rho}(\alpha,-)\geq const=\frac{1}{2}\max_{\eta',\eta''\in A}\rho(\eta',\eta'')\ \forall\ \alpha\in A$$

**Sulimova V.V., Seredin O.S., Mottl V.V. Metrics on the basis of optimal alignment of biomolecular sequences // JMLDA, 2016 (In Russian)*

CONDITIONAL DISSIMILARITY MEASURES OF BIOMOLECULAR SEQUENCES

$\mathbf{w} \in W_{N'N''}$ - permissible alignment

In terms of initial sequences:

$$r_1(\omega', \omega'' \mid \mathbf{w}) = \sum_{i:\mathbf{w}_{i,1}\neq 0, \mathbf{w}_{i,2}\neq 0} \rho(\alpha'_{\mathbf{w}_{i,1}}, \alpha''_{\mathbf{w}_{i,2}}) + \sum_{i:\mathbf{w}_{i,1}=0 \ or \ \mathbf{w}_{i,2}=0} \beta$$

$$r_2(\omega', \omega'' \mid \mathbf{w}) = \sqrt{\sum_{i:\mathbf{w}_{i,1}\neq 0, \mathbf{w}_{i,2}\neq 0} \rho^2(\alpha'_{\mathbf{w}_{i,1}}, \alpha''_{\mathbf{w}_{i,2}}) + \sum_{i:\mathbf{w}_{i,1}=0 \ or \ \mathbf{w}_{i,2}=0} \beta^2}$$

$\beta = \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \ \forall \alpha \in A$ - gap penalty

In terms of extended sequences:

$$r_1(\omega', \omega'' \mid \mathbf{w}) = \sum_{i=1}^{N_{\mathbf{w}}} \tilde{\rho}(\tilde{\alpha}'_{\mathbf{w},i}, \tilde{\alpha}''_{\mathbf{w},i})$$

$$r_2(\omega', \omega'' \mid \mathbf{w}) = \sqrt{\sum_{i=1}^{N_{\mathbf{w}}} \tilde{\rho}^2(\tilde{\alpha}'_{\mathbf{w},i}, \tilde{\alpha}''_{\mathbf{w},i})}$$

METRICS ON THE SET OF BIOMOLECULAR SEQUENCES

$$r_1(\omega',\omega'') = \min_{\mathbf{w}\in W_{N'N''}} r_1(\omega',\omega''\,|\,\mathbf{w}) = \min_{\mathbf{w}\in W_{N'N''}} \sum_{i=1}^{N_{\mathbf{w}}} \tilde{\rho}(\tilde{\alpha}'_{\mathbf{w},i},\tilde{\alpha}''_{\mathbf{w},i})$$

$$r_2(\omega',\omega'') = \sqrt{\min_{\mathbf{w}\in W_{N'N''}} r_2(\omega',\omega''\,|\,\mathbf{w})} = \sqrt{\min_{\mathbf{w}\in W_{N'N''}} \sum_{i=1}^{N_{\mathbf{w}}} \tilde{\rho}^2(\tilde{\alpha}'_{\mathbf{w},i},\tilde{\alpha}''_{\mathbf{w},i})}$$

Theorem 3*.

For any metric on the extended set of elements

$$\tilde{\rho}(\alpha',\alpha''),\quad \tilde{\alpha}',\tilde{\alpha}'' \in \tilde{A} \qquad \tilde{A} = A \cup \{-\} = \{\alpha^1,\ldots,\alpha^m,-\} = \{\tilde{\alpha}^1,\ldots,\tilde{\alpha}^m,\tilde{\alpha}^{m+1}\}$$

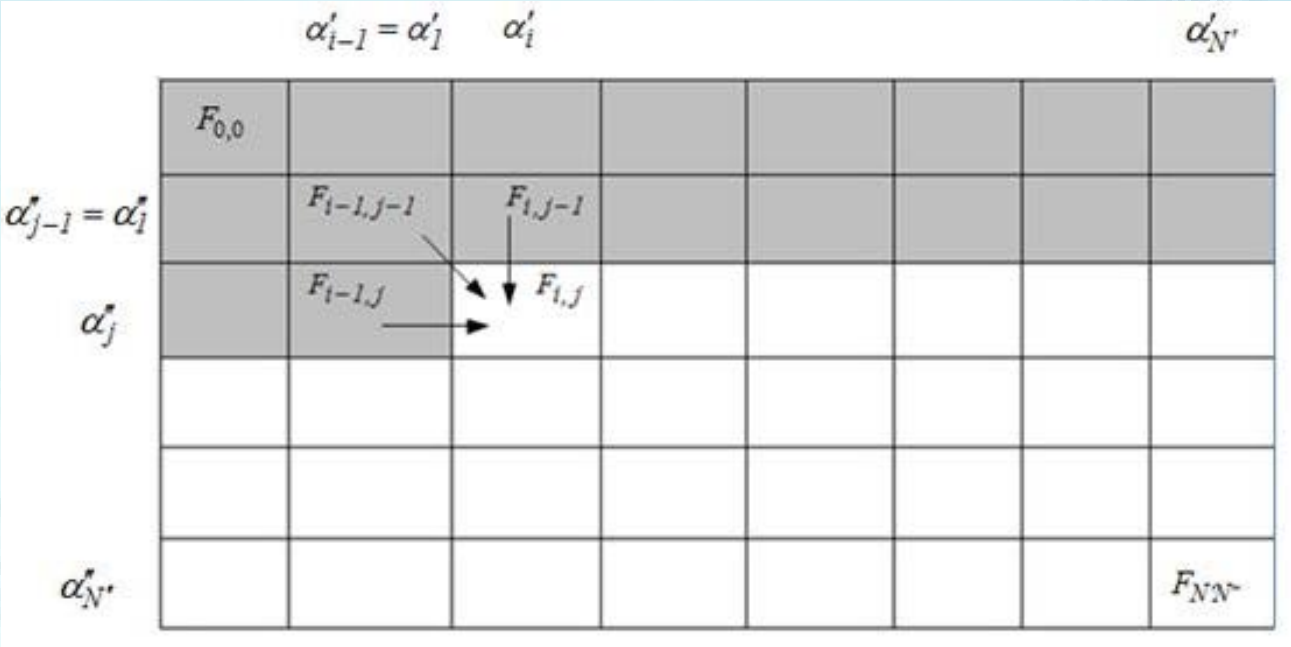
functions $r_1(\omega',\omega'')$ and $r_2(\omega',\omega'')$ are metrics on the set of sequences

**Sulimova V.V., Seredin O.S., Mottl V.V. Metrics on the basis of optimal alignment of biomolecular sequences // JMLDA, 2016 (In Russian)*

COMPARING THE PROPOSED METRIC WITH THE TRADITIONAL NEEDLEMAN-WUNSCH ALIGNMENT

THE PROPOSED METRIC	NEEDLEMAN-WUNSCH ALIGNMENT
Comparing of elements	
metric $\rho(\alpha',\alpha'')$	similarity measure $s(\alpha',\alpha'')$
The criterion	
$\min\limits_{\mathbf{w}}\left(\sum\limits_{\substack{i:\mathbf{w}_{i,1}\neq 0,\\ \mathbf{w}_{i,2}\neq 0}}\rho(\alpha'_{\mathbf{w}_{i,1}},\alpha''_{\mathbf{w}_{i,2}})+\sum\limits_{\substack{i:\mathbf{w}_{i,1}=0\\ or\ \mathbf{w}_{i,2}=0}}\beta\right)$	$\max\limits_{\mathbf{w}}\left(\sum\limits_{\substack{i:\mathbf{w}_{i,1}\neq 0,\\ \mathbf{w}_{i,2}\neq 0}}s(\alpha'_{\mathbf{w}_{i,1}},\alpha''_{\mathbf{w}_{i,2}})+\sum\limits_{\substack{i:\mathbf{w}_{i,1}=0\\ or\ \mathbf{w}_{i,2}=0}}\beta\right)$
Gap penalty	
$\beta \geq 0$	$\beta \leq 0$
Type:	
METRIC	SIMILARITY MEASURE

ALGORITHM OF COMPUTING THE PROPOSED METRIC



Initializing: $F_{0,0} = 0, \quad F_{i,0} = i\beta, \; i = 1,\ldots,N'', \quad F_{0,j} = j\beta, \; j = 1,\ldots,N''$

Finding partial criterion values for each $i = 1,\ldots,N'', \; j = 1,\ldots,N''$:

$$F_{i,j} = \min \begin{cases} F_{i-1,j-1} + \rho(\alpha'_i, \alpha''_j), \\ F_{i-1,j} + \beta, \\ F_{i,j-1} + \beta, \end{cases} \qquad F_{i,j} = \min \begin{cases} F_{i-1,j-1} + \rho^2(\alpha'_i, \alpha''_j), \\ F_{i-1,j} + \beta^2, \\ F_{i,j-1} + \beta^2, \end{cases}$$

Finding partition criterion value:

$$r_1(\omega', \omega'') = F_{N'N''} \qquad r_2(\omega', \omega'') = \sqrt{F_{N'N''}}$$

EXAMPLES OF OPTIMAL ALIGNMENTS

Needleman-Wunsch algorithm
with PAM250 substitution matrix
and default penalty value

```
Identities = 14/33 (42%), Positives = 16/33 (48%)
AN--AK-N---I-TC--A-EEPAM-HNLACA-T
||  || :   |  |  |  :|||    ||| |
ANCCAKTSLKVILCCIMAPPDPAMLAAAACAST
```

```
Identities = 48/214 (22%), Positives = 99/214 (46%)
001  MA-PAVARACAW---A-L---LA-A--LL-WLPPA-GATRRA--D-SAESILAERCRG-NL--L
     |  :::|: ::   : :   :: |  || ::  :  |: :|  : | :::     |: |:  :
001  MIFEVMSRTFVFNRRGNMKVSMSVACILLVYVYRVFFASAEAPMNWSDRDLTYGFVRAPNVTTI

046  L-AD-RPQHE-EA---A-P-G--L-AGIPIRGRCSPPEAALWYEDTGETYWANPYAVARGLAED
     :  :  |  :  :   | | :  : :||:|: :|| ||  |||| :| : |:||   :  | ||
065  VHLECIPTSKLTSMRYAEPSSDEVPSGIIIKTNCSLPEPILWYERVGVAAWVNPIIGTSLLLED

099  IRRVLADTPVYRDLAIQVLNSAPGLP--H-EVR-A---PLPPPPRG-CV--LP-PRY-HT----
     : |   | |   ::  : : |: :|  | : | |    |    | |   :  |: ::
129  VLRSRLDDSVKAGIGTLLSKIAYLIPTSHLRNRGAGCINLYASHDGTCYGSVHFDRFERSADDD

147  T-GP-C-G--P---GDGMYR--
     : |  | :  |   : |  |
193  NRGSFCRNKTPRLRNGGRPRET
```

the proposed method $r_1(\omega', \omega'')$
with PAM250-based metric
and the penalty

$$\beta = \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'')$$

```
Identities = 13/33 (39%), Positives = 16/33 (48%)
AN--AK-N--IT--C--A-EEPAM-HNLACA-T
||  || :  :   |  |  :|||    ||| |
ANCCAKTSLKVILCCIMAPPDPAMLAAAACAST
```

```
Identities = 60/225 (27%), Positives = 96/225 (43%)
001  -----MAPA-V-A-R-AC--AWALLA-ALL----WL--PPAGA----T-R-------RADSAES
          |: : |  |     : : :|  ||     :   | |   : |       || :: :
001  MIFEVMSRTFVFNRRGNMKVSMS-VACILLVYVYRVFFASAEAPMNWSDRDLTYGFVRAPNVTI

035  I--LAERCRGNL-LLADR---PQHEEAAPGLAGIPIRGRCSPPEAALWYEDTGETYWANPYAVA
     |  | | |  :  | : |   |: :| :|  :||:|: :|| ||  |||| :| : |:||  ::
064  IVHL-E-CIPTSKLTSMRYAEPSSDE-VP--SGIIIKTNCSLPEPILWYERVGVAAWVNP-IIG

093  RG-LAEDI-RRVLADTPVYRDLAI-QVLN-SAPGL-P--H-EVR-A---PLPPPPRG-C---VL
      : | ||: |  | |: | :  :|  :|:  |: | |  | : | |    |     | |  |
122  TSLLLEDVLRSRLDDS-V-KA-GIGTLLSKIAY-LIPTSHLRNRGAGCINLYASHDGTCYGSVH

141  -P--PR--YHTT-GP-C-G--P--GD-GMYR--
          |    : |  | :  |   : |  |
182  FDRFERSADDDNRGSFCRNKTPRLRNGGRPRET
```

DATA FOR EXPERIMENTS

Amino acid sequences of herpes simplex virus
from VIDA (Virus Database at University College London)

	Description	Homologous protein families (HPF)	Number of proteins
Class 1 (109 proteins)	Glycoprotein H	12	52
		42	39
		531	18
Class 2 (77 proteins)	Glycoprotein L	47	30
		50	32
		114	13
		296	2
Class 3 (48 proteins)	Glycoprotein M	20	48

EXPERIMENTAL DESIGN

Basic ways to compare sequences:

1. Needleman-Wunsch similarity measure $S_1(\omega', \omega'')$
2. Smith-Waterman similarity measure $S_2(\omega', \omega'')$
3. The proposed metric $r(\omega', \omega'')$

For using SVM:

(1) and (2):

Kernels in secondary features space

$$K_i(\omega', \omega'') = \left[k_{lt} = (S_i^{<l>})^T S_i^{<t>} \right], \quad l, t = 1, \ldots, N], \quad i = 1, 2$$

(3) : radial basis kernel

$$K_3(\omega', \omega'') = \exp\left(-\alpha r^2(\omega', \omega'')\right) \quad \text{with} \quad \alpha = 0.01$$

34 recognition tasks:

- one-against-all recognition for classes (3 tasks)
- one-against-all recognition for HPFs (7 tasks)
- one-against-one recognition for classes (3 tasks)
- one-against-one recognition for HPFs (21 tasks)

EXPERIMENTAL RESULTS

LOO-error percentages for one-to-all recognition

Class	NW	SW	Metric
hpf 12	15,0215	15,0215	**14,5923**
hpf 20	0,4292	**0**	**0**
hpf 42	**0**	0,4292	0,4292
hpf 47	4,721	**0**	**0**
hpf 50	0,4292	**0**	**0**
hpf 114	4,721	0,8584	**0,4292**
hpf 531	**15,0125**	**15,0125**	18,4549
class 1	0,8584	**0,4292**	**0,4292**
class 2	0,8584	**0,4292**	**0,4292**
class 3	0,4292	**0**	**0**

LOO-error percentages for one-to-all recognition

Task	NW	SW	Metric
class 2 vs class 3	12,3256	**0**	**0**
hpf 42 vs hpf 47	0,4292	**0**	**0**
hpf 42 vs hpf 114	**0**	1,9231	**0**
hpf 47 vs hpf 114	2,3256	**0**	**0**
hpf 531 vs hpf 12	**48,5714**	51,4286	50,000
hpf 531 vs hpf 42	**1,7544**	3,5088	**1,7544**

THANK YOU FOR YOUR ATTENTION!