Improving the Needleman-Wunsch algorithm with the DynaMine predictor

Olivier Boes

Advisors: **Tom Lenaerts, Wim Vranken, Elisa Cilia.**

Université Libre de Bruxelles — September 2014

Reminder on sequence alignments

- A **protein sequence alignment** is something like this:

```
MSDINATRLPAWLVDC-PCVGDDINRLLTRGENSLC   (Amanita virosa)
MSDINATRLPAWLVDC-PCVGDDVNRLLTRGE-SLC   (Amanita bisporigera)
MSDINATRLPIWGIGCDPCIGDDVTALLTRGEASLC   (Amanita phalloides)
----------IWGIGCNPCVGDEVTALLTRGEA---   (Amanita fuligineoides)
```

  It tries to identify regions of similarity between different proteins believed to be related (e.g. common ancestor).

- **Applications**: sequence identification, homology modeling, genome assembly, motif discovery, phylogenetics,...

- In this thesis, we focus on **pairwise global alignments**:
  - only two protein sequences are aligned,
  - all amino acid residues are aligned.

What does the thesis title mean?

- **Needleman-Wunsch** is a sequence alignment algorithm.
  It aligns proteins using their amino acid sequences alone.

- **DynaMine** is a predictor of protein backbone flexibility.
  It gives us some information on a protein structure.

- **Structure is more conserved than sequence.**
  Therefore we want to create a Needleman-Wunsch variant
  which uses the structural information provided by DynaMine.

Could such a variant produce better alignments?

This question is central to the thesis.

Outline of what was done

Basically:

1. Choosing datasets of reference alignments.
2. Creating DynaMine-based score matrices.
3. Using them in our Needleman-Wunsch variant.
4. Comparing computed and reference alignments.
5. Results, discussion, conclusion.

Lots of programming (mostly C and Python) was required!

The BAliBASE benchmark database

Contains multiple sequence alignments believed to be correct.

Five BAliBASE datasets were used:

- **RV11** and **RV12**: sequences with low residue identity.
- **RV20**: families aligned with a highly divergent sequence.
- **RV30**: alignments of divergent protein subfamilies
- **RV50**: sequences with large internal insertions

Each one is partitioned into a **training** set and a **test** set.

```
...GXVETDD--------------------GRSFVXADLPGLIEGA-HQGVGLGHQ-FLRHIERTRVIVHVIDXSGL-------EGRDPYDDY...
...ADAEIRRCPNCGRYSTSPVCPYCGHETEFVRRVSFIDAPGHEALMTTMLAGASLM---------DGAILVIAANEP--------CPRPQTRE...
...WKFETP---------------------KYQVTVIDAPGHRDFIKNMITGTSQA---------DCAILIIAGGVGEFEAG--ISKDGQTRE...
...VEYETA---------------------KRHYSHVDCPGHADYIKNMITGAAQM---------DGAILVVSAADG---------PMPQTRE...
...GATEIPXDVIEGICGDF---LKKFSIRETLPGLFFIDTPG--AFTTLRKRGGALA---------DLAILIVDINEG---------FKPQTQE...
...LGAYTD---------------------DLDYVFYDVLGDVVCGGFAMPIREG---------KAQEIYIVASGEMMALYA--ANNISKGIQ...
...GIIETQFSFK-----------------DLNFRMFDVGGQRSERKKWIHCFEG----------VTCIIFIAALSAYDMVLVEDDEVNRMHE...
```

Reference: Julie D. Thompson, Patrice Koehl, Raymond Ripp, Olivier Poch. *BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark* Proteins: Structure, Function, and Bioinformatics, *61(1):127–136*, 2005.

The DynaMine flexibility predictor

Predicts protein backbone flexibility at the residue-level.

amino acid sequence

$(x_1 \quad x_2 \quad x_3 \quad \cdots \quad x_m)$

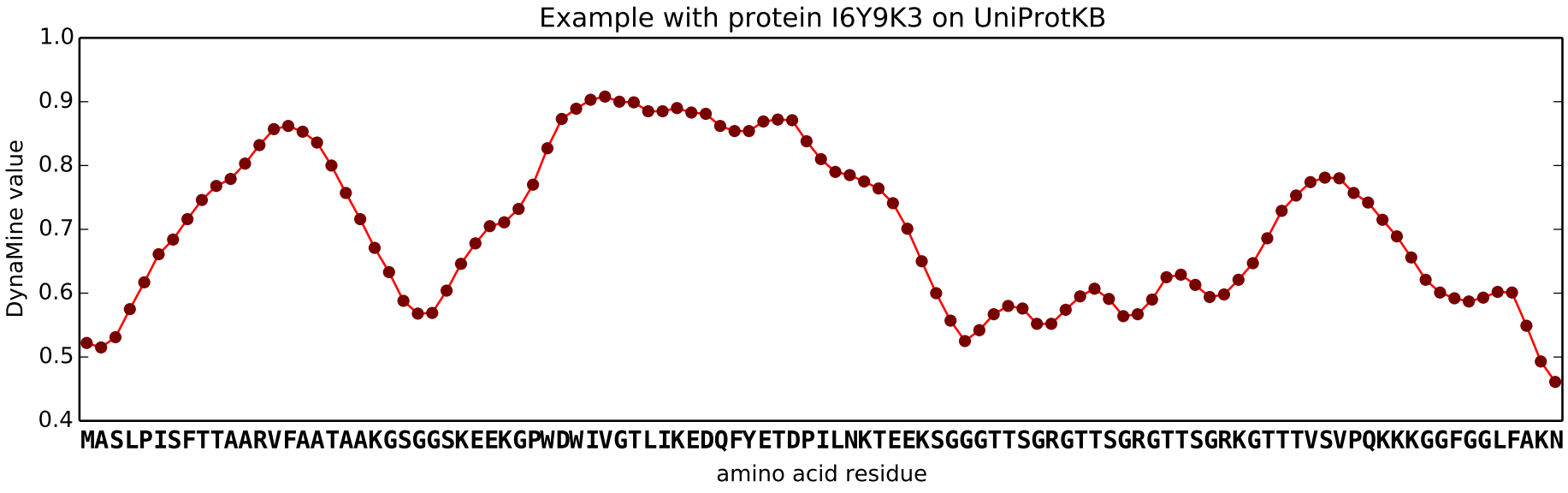
$x_i \in \{\text{ARNDCQEGHILKMFPSTWYV}\}$

DynaMine $\longmapsto$

flexibility value sequence

$(u_1 \quad u_2 \quad u_3 \quad \cdots \quad u_m)$

more flexible $0 \leq u_i \leq 1$ more rigid



Reference: Elisa Cilia, Rita Pancsa, Peter Tompa, Tom Lenaerts, Wim F. Vranken.
From protein sequence to dynamics and disorder with Dynamine.
Nature Communications, *4:2741*, 2013.

The Needleman-Wunsch variant

Algorithm for aligning two sequences $(x_1 \cdots x_m)$ and $(y_1 \cdots y_n)$.

In its most generalized version, it requires:

- substitution scores $\mathrm{sub}(i, j)$ for aligning x_i with y_j
- opening and extending gap penalties (not necessarily constant)

Usually: $\mathrm{sub}(i,j) := \mathrm{seqS}(x_i, y_j)$

Variant: $\mathrm{sub}(i,j) := \alpha \cdot \mathrm{seqS}(x_i, y_j) + (1-\alpha) \cdot \mathrm{dynS}(u_i, v_j)$

Several dynS matrices were created using BLOSUM and BAliBASE.

Custom Needleman-Wunsch alignment software was also developed.

Implementation of the NW algorithm: C source code available on `https://github.com/oboes/gotoh`

BLOSUM matrices: how they are created

1. Choose a reference dataset of blocks (gap-free alignments).
2. Cluster together sequences with more than T% similarity.
3. Compute log-odds scores (i.e. log-likehood ratios).

$$\mathrm{BLOSUM}T(x,y) := \frac{1}{\lambda}\log\left(\frac{\mathrm{P}(\text{substitution } x \leftrightarrow y)}{\mathrm{P}(\text{residue } x)\cdot\mathrm{P}(\text{residue } y)}\right)$$

BLOSUM62 created with my script

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0
R	-2	5	0	-2	-3	1	0	-2	0	-3	-2	2	-2	-3	-2	-1	-1	-3	-2	-2
N	-1	0	6	1	-3	0	0	-1	1	-3	-3	0	-2	-3	-2	0	0	-3	-2	-3
D	-2	-2	1	6	-4	0	2	-2	-1	-3	-3	-1	-3	-4	-2	0	-1	-4	-3	-3
C	-1	-3	-3	-4	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-3	-2	-1
Q	-1	1	0	0	-3	5	2	-2	1	-3	-2	1	0	-3	-1	0	0	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	-1	-2	-3	-3	-3
H	-2	0	1	-1	-3	1	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-2	-1	-2
L	-2	-2	-3	-3	-1	-2	-3	-4	-3	2	4	-2	2	1	-3	-2	-2	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-2	-2	-3	-1	0	-2	-3	-1	1	2	-1	6	0	-2	-1	-1	-2	-1	0
F	-2	-3	-3	-4	-2	-3	-3	-3	-2	0	1	-3	0	6	-3	-2	-2	1	3	-1
P	-1	-2	-2	-2	-3	-1	-1	-2	-2	-3	-3	-1	-2	-3	7	-1	-1	-4	-3	-2
S	1	-1	0	0	-1	0	0	-1	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	0	-1	-2	-2	-1	-2	-1	-1	-2	-1	1	5	-3	-2	0
W	-3	-3	-3	-4	-3	-2	-3	-3	-1	-2	-2	-3	-2	1	-4	-3	-3	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	1	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-2	-3	-3	-1	-2	-2	-3	-3	2	1	-2	0	-1	-2	-2	0	-3	-1	4

BLOSUM62 used in most softwares

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLOSUM62 claimed to be correct

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0
R	-2	5	0	-2	-3	1	0	-2	0	-3	-2	2	-2	-3	-2	-1	-1	-3	-2	-3
N	-1	0	6	1	-3	0	0	-1	1	-3	-3	0	-2	-3	-2	0	0	-3	-2	-3
D	-2	-2	1	6	-3	0	2	-2	-1	-3	-3	-1	-3	-3	-2	0	-1	-4	-3	-3
C	-1	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	1	-3	-2	1	0	-3	-1	0	0	-2	-2	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	-1	-2	-2	-3	-3
H	-2	0	1	-1	-2	1	0	-2	7	-3	-3	-1	-1	-2	-1	-2	-1	-2	1	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-2	-1	2
L	-2	-2	-3	-3	-1	-2	-3	-4	-3	2	4	-2	2	1	-3	-2	-1	-1	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-2	-2	-3	-1	0	-2	-3	-1	1	2	-1	6	0	-2	-1	-1	-2	-1	0
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	1	-3	0	6	-3	-2	-2	1	3	-1
P	-1	-2	-2	-2	-3	-1	-1	-2	-2	-3	-3	-1	-2	-3	7	-1	-1	-3	-3	-2
S	1	-1	0	0	-1	0	0	-1	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	0	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
W	-3	-2	-3	-4	-3	-2	-3	-2	-1	-2	-1	-3	-2	1	-3	-3	-3	11	2	-3
Y	-2	-2	-2	-3	-2	-2	-2	-3	1	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	2	1	-2	0	-1	-2	-2	0	-3	-1	4

Reference: Mark P. Styczynski, Kyle L. Jensen, Isidore Rigoutsos, Gregory Stephanopoulos. *BLOSUM62 miscalculations improve search performance*. Nature Biotechnology, *26:274–275*, 2008.

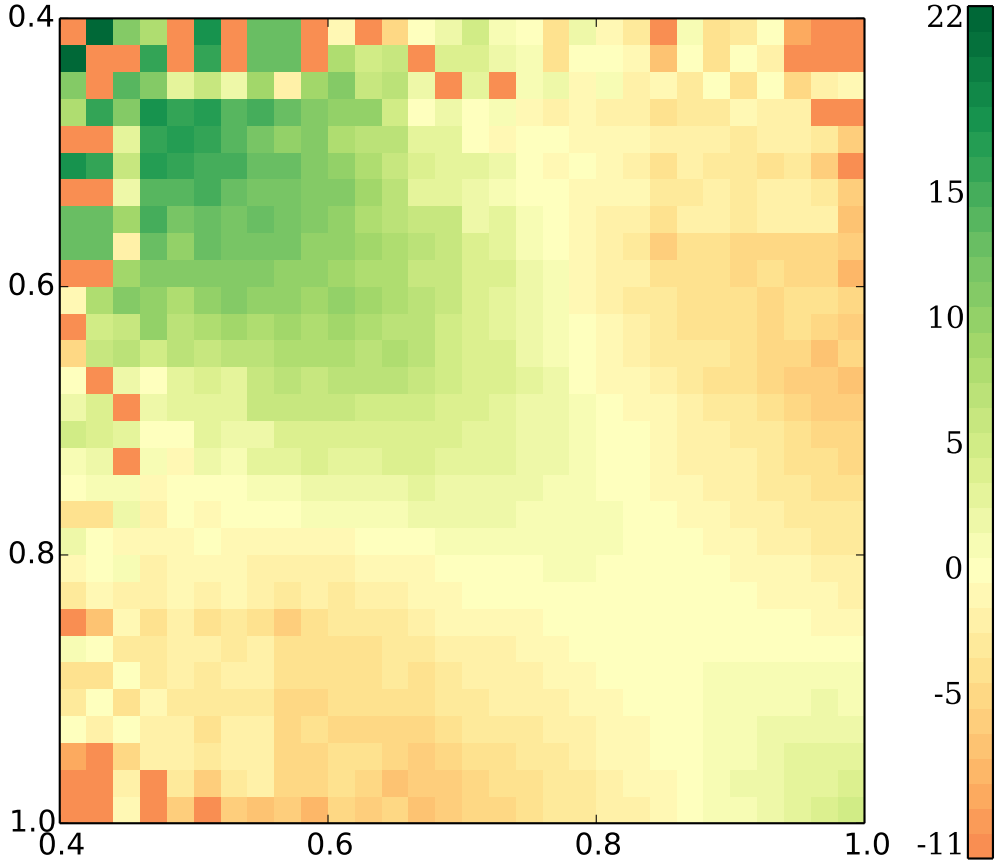
BLOSUM matrices: with DynaMine values

- DynaMine values converted to integers using 50 bins
- Blocks taken from BAliBASE alignments of DynaMine values
- Same expected score for $\mathrm{seqBLOSUM}$ and $\mathrm{dynBLOSUM}$
- Each BAliBASE dataset has its own $\mathrm{dynBLOSUM}$ matrix

Example:

The $\mathrm{dynBLOSUM62}$ matrix created from the BAliBASE `RV30` training dataset

(there are no values < 0.4)



Summary of how we do our experiments

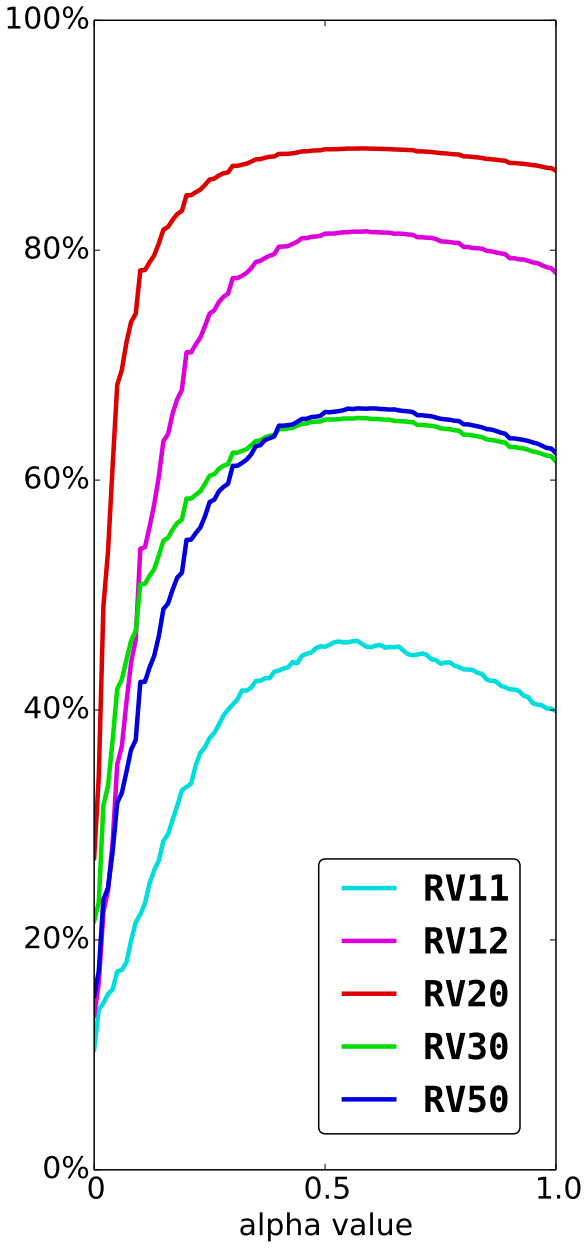
substitution scores:

$$\mathrm{sub}(i,j) \;:=\; \alpha \cdot \mathrm{seqBLOSUM62}(x_i, y_j) \;+\; (1-\alpha) \cdot \mathrm{dynBLOSUM62}(u_i, v_j)$$

- Clustering threshold: 62%
- Gap penalties: 10 for opening, 0.5 for extending
- Matrix from **RVxy** training set used for aligning **RVxy** test set
- Quality measure: $\frac{\#\text{ pairs correctly aligned in } \texttt{RVxy}}{\#\text{ pairs aligned in reference } \texttt{RVxy}}$ (sum-of-pairs score)
- Computational cost?
  - α goes from 0 to 1 by increments of 0.01
  - $80\,000$ pairs of sequences to align for each α

  $\Longrightarrow$ more than 8 million alignments to compute

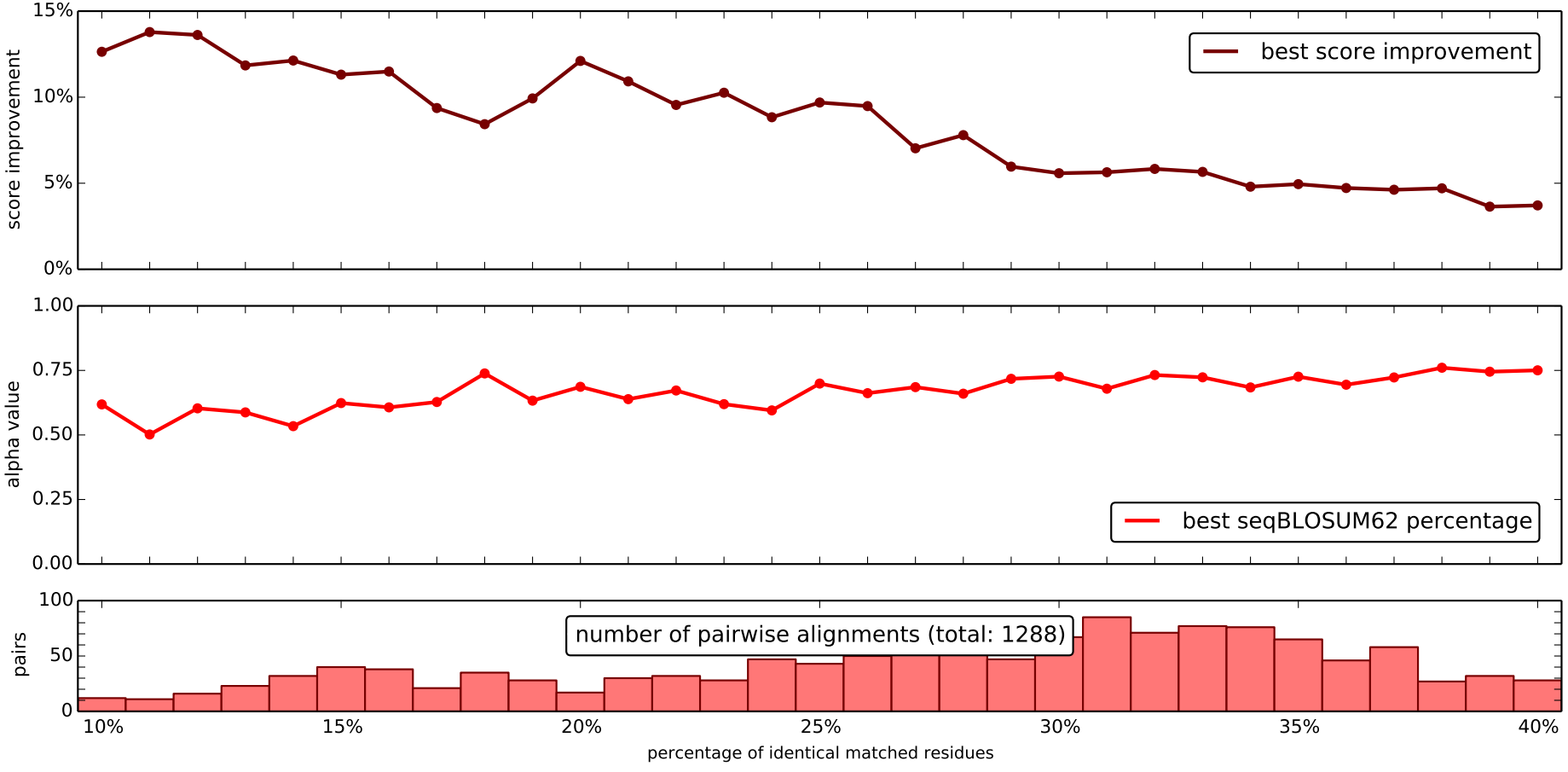
Results for each BAliBASE dataset



dataset	RV11	RV12	RV20	RV30	RV50
multiple alignments	19	22	21	15	8
pairwise alignments	412	1655	23552	50899	3429
score without DynaMine	40%	78%	87%	62%	62%
best score with DynaMine	46%	82%	89%	65%	66%
possible score increase	**6%**	**4%**	**2%**	**3%**	**4%**
α *producing best score*	**0.57**	**0.59**	**0.58**	**0.58**	**0.57**

Results for dissimilar sequences

- Pairwise alignments from **RV11** and **RV12** with ≥ 150 pairs.
- Datasets for each residue identity percentage are created.
- Possible score increase and corresponding α determined.



Example of an improved pairwise alignment

reference **RV11** alignment: 145 aligned pairs and 11% residue identity:

```
RKNLVQFGVGEKNGSVRWVMNALGVKDDWLLVPSHAYKFEKDYEMMEFYFNRGGTYYSISAGNVVIQSLDVGFQDVVLMKVPTIPKFRDITQHFIKKGDVPRA-LNR...
LEADRLFDVKNEDGDVIGHALAMEG---KVMKPLHVK---------GTIDHP-----VLSKLKFTKSSA----YDMEFAQLP--------VNMRSEAFTYTSEH--P...

...LATLVTTVNGTPMLISEGPLKMEEKATYVHKKNDGTTVDLTVDQAWRGKGEGLPGMCGGALVSSNQSIQNAILGIHVAGGN---SILVAKLVT-QEMFQNIDKKI
...EGFYNWH------HGAVQYSGG---------------RFTIPRGV------GGRGDAGRPIMDN----SGRVVAIVLGGADEGTRTALSVVTWNSKGKTIKTTPE
```

aligned without DynaMine ($\alpha = 1.00$): $27/145 = \mathbf{19}\%$ correct pairs:

```
RKNLVQFGVGEKNGSVRWVMNALGVKDDWLLVPSHAYKFEKDYEMM-EFYFNRGGTYYSISAGNVVIQSLDVGFQDVVLMKVPTIPK-----FRDITQHFIKKGDVP...
LEADRLFDVKNEDGDV--IGHALAMEGK-VMKPLHV-KGTIDHPVLSKLKFTKSSAY------DMEFAQLPVNMRSEAFTYTSEHPEGFYNWHHGAVQYSGGRFTIP...

...RALNRLATLVTTVNGTPMLISEGPLKMEEKATYVHKKNDGTTVDLTVDQAWRGKGEGLPGMCGGALVSSNQSIQNAILGIHVAGGNSILVAKLVTQEMFQNIDKKI
...RGVGGRGDA-----GRPIMDNSGRVV----AIVLGGADEGTRTALSV-VTWNSKGKTIK----------------------------------------TTPE
```

aligned with DynaMine ($\alpha = 0.57$): $57/145 = \mathbf{39}\%$ correct pairs:

```
RKNLVQFGVGEKNGSVRWVMNALGVKDDWLLVPSHAYKFEKDYEMMEFYFNRGGTYYSISAGNVVIQSLDVGFQDVVLMKVPTIPKFRDITQHFIKKGDVPRALNRL...
LEADRLFDVKNEDGDVIGHALAMEGK---VMKPLHVKGTIDHPVLSKLKFTKSSAY-------------------------------------------------...

...ATLVTTVNGTPMLISEGPLKMEEKA-TYVHKKNDGT------TVDLTVDQAWRGKGEGLPGMCGGALVSSNQSIQNAILGIHVAGGNSILVAKLVTQEMFQNIDKKI
...----------DMEFAQLPVNMRSEAFTYTSEHPEGFYNWHHGAVQYSGGRFTIPRGVGGRGDAGRPIMDNSGRVVAIVLGGADEGTRTALSVVTWNSKGKTIKTTPE
```

Pairwise alignment: sequences '1hav_A' and '1svp_A' from the BB11011 file in the **RV11** BAliBASE dataset.

Conclusion

Good:

- Better alignments produced, and α value stays the same across datasets.
- Best results obtained with dissimilar sequences.

Bad:

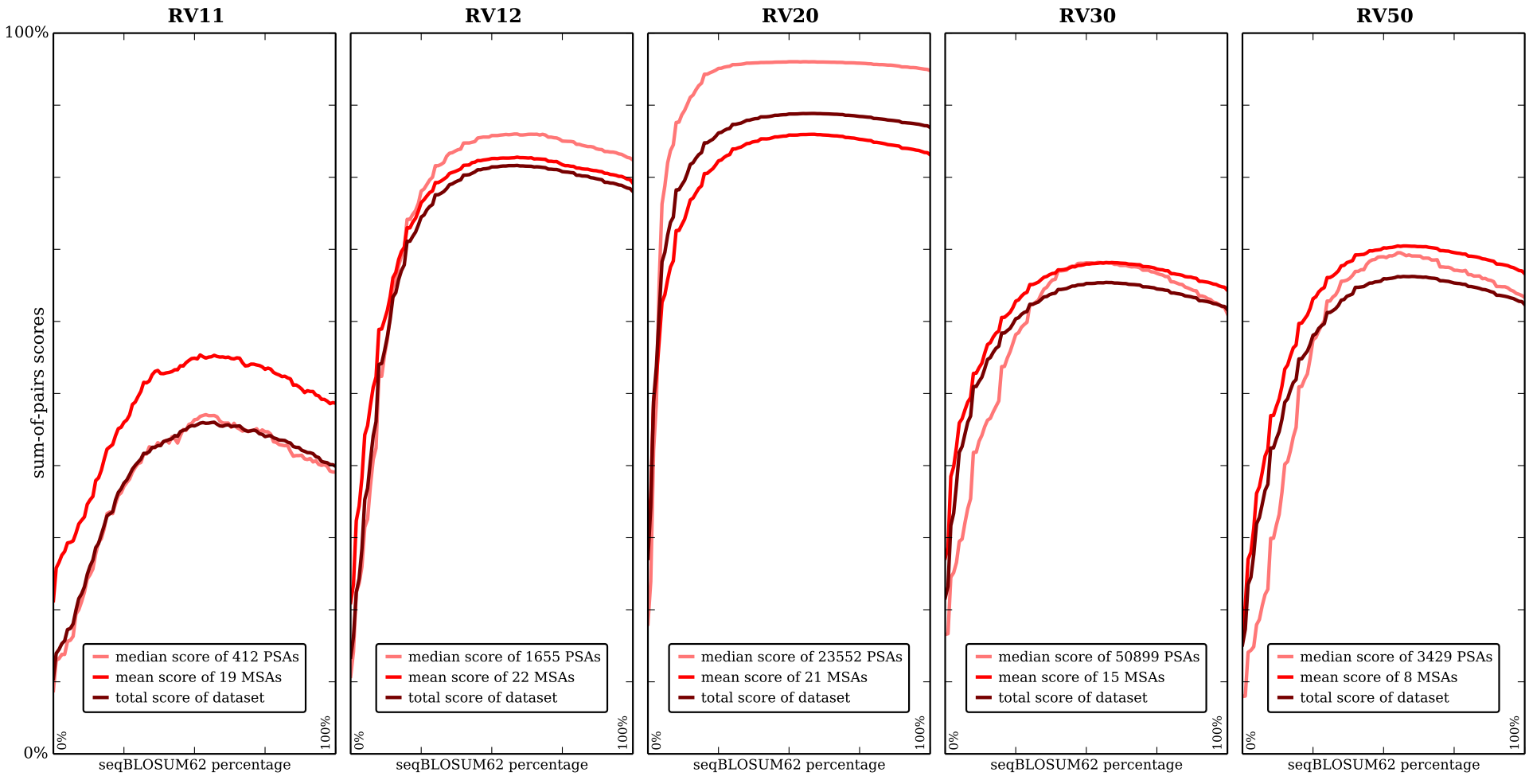
- Improvements are very small, and largest ones occur in smallest datasets.
- Datasets of dissimilar sequences are too small.

What could be interesting to try:

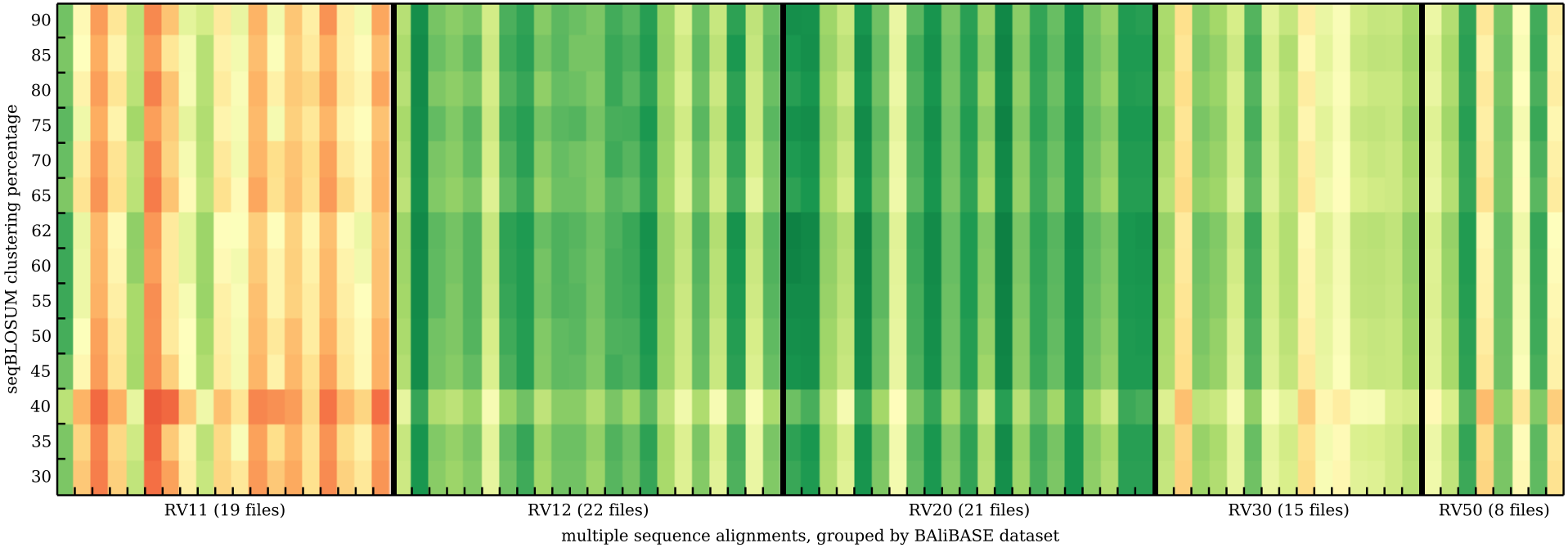
- Alignment benchmarks for dissimilar or disordered proteins.
- Using DynaMine differently for creating substitution scores.
- Gap penalties also depending on DynaMine values.
- Aligning with something else than Needleman-Wunsch.
- Measuring quality with something else than sum-of-pairs.

Questions?

Appendix: different sum-of-pairs scores



Appendix: best seqBLOSUM matrices



	30	35	40	45	50	55	60	62	65	70	75	80	85	90
RV11	35	38	36	39	35	34	41	40	39	37	35	26	32	31
RV12	75	77	75	77	75	74	79	78	77	76	75	65	72	71
RV20	85	86	85	86	85	84	87	87	87	86	85	79	84	83
RV30	59	60	59	61	59	58	62	62	61	60	59	51	57	56
RV50	59	61	59	61	60	58	63	62	62	61	60	52	58	57

Appendix: DynaMine bias near end sequence ends