Improving Multiple Sequence Alignments with Constraint Programming and Local Search

Marco Correia, **Fábio Madeira**, Pedro Barahona and Ludwig Krippahl

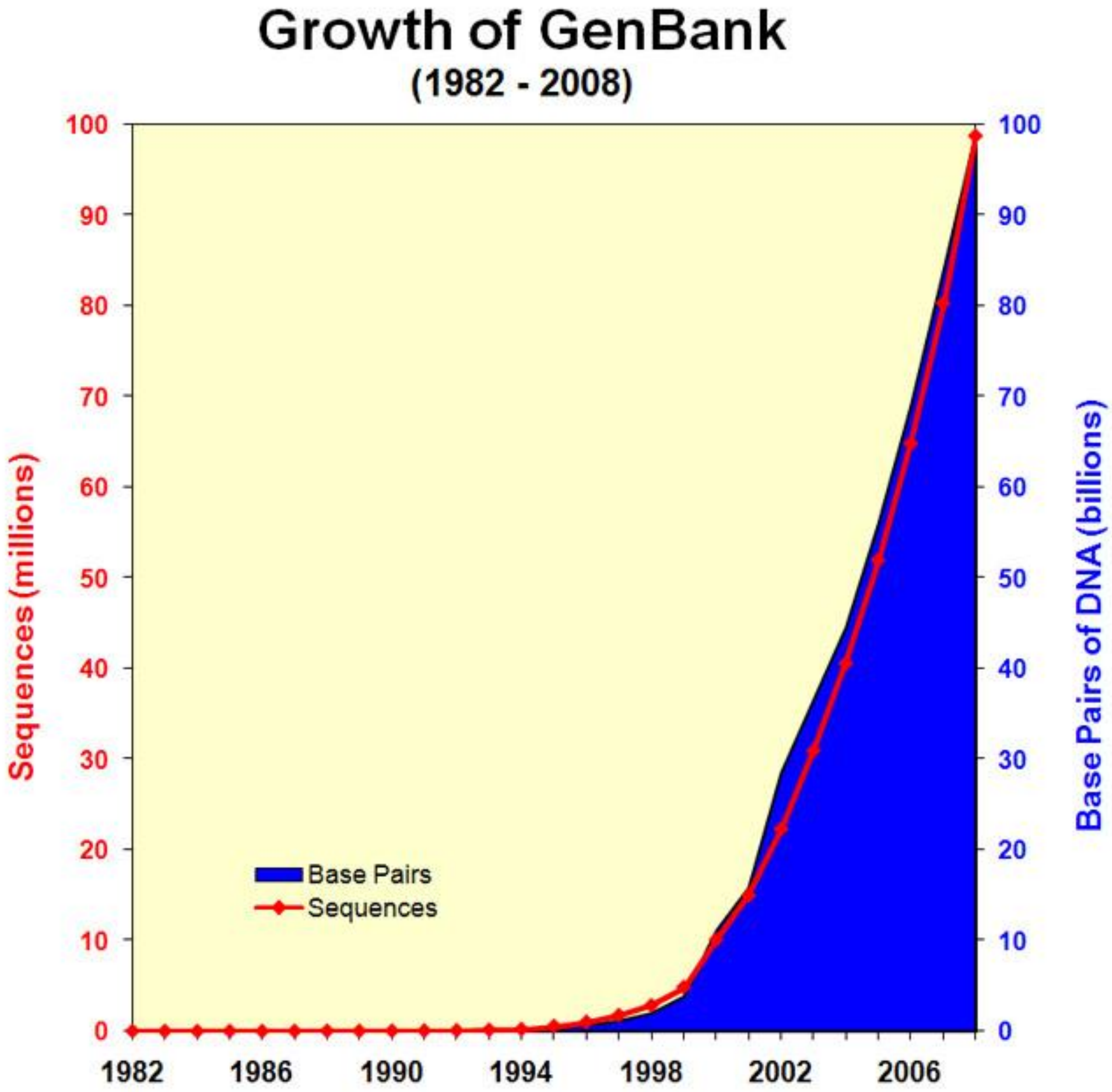
CENTRIA-DI, FCT/UNL, Portugal

WCB11 - Workshop on Constraint Based Methods for Bioinformatics - 12 September 2011

OUTLINE

1. Introduction
   1. Our proposal
2. Method
   1. CP Model
   2. Search
   3. Local Search
   4. Experiments
3. Results and Discussion
4. Conclusions

INTRODUCTION



http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html, September 2011

INTRODUCTION

```
S1      QFGLFSPEEIRASSVALIRTPYPETLENGVPKESGLVCAGHFGHIELVK

S2      QFGLFSPEEIKRMSVVHVEYPETMDEQRQRPRTKGLECPGHFGHIELAT

S3      ELGVLDPEIIKKISVCEIVPNVDIYKDGRFPREGGLYCPGHFGHIELAK
```

Protein sequences

```
S1      QFGLFSPEEIRASSVALIRTPYPETLENG--VPKESGLVCAGHFGHIELVK
S2      QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
```

Pairwise alignment

```
S2      QFGLFSPEEIKRMSVVHVE-YPETMDEQRQRPRTKGLECPGHFGHIELAT
S3      ELGVLDPEIIKKISVCEIVPNVDIYKDGR-FPREGGLYCPGHFGHIELAK
```

Pairwise alignment

INTRODUCTION

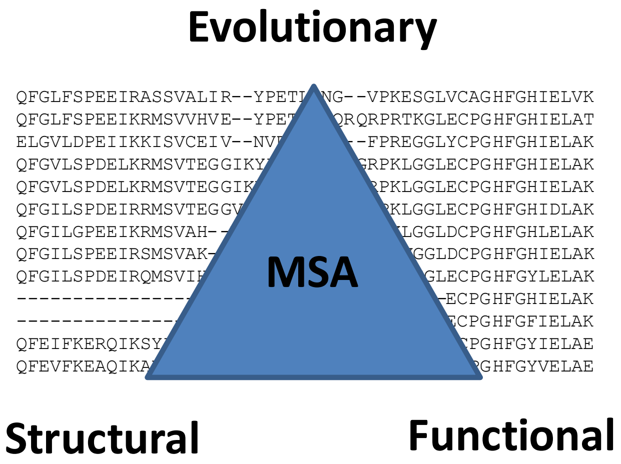
```
S1    QFGLFSPEEIRASSVALIRTPYPETLENGVPKESGLVCAGHFGHIELVK
S2    QFGLFSPEEIKRMSVVHVEYPETMDEQRQRPRTKGLECPGHFGHIELAT
S3    ELGVLDPEIIKKISVCEIVPNVDIYKDGRFPREGGLYCPGHFGHIELAK
⋮
```

Protein sequences

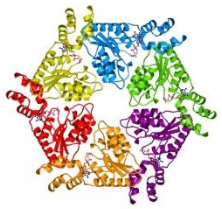
```
S1    QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAGHFGHIELVK
S2    QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
S3    ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
S4    QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
S5    QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
S6    QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
S7    QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
S8    QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
S9    QFGILSPDEIRQMSVIH----VEHSETTEKGKPKVGGLECPGHFGYLELAK
S10   --------------------------------------ECPGHFGHIELAK
S11   --------------------------------------ECPGHFGFIELAK
S12   QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
S13   QFEVFKEAQIKAYAKCIIEHAKSYEHG----QPVRGGIECPGHFGYVELAE
```

Multiple Sequence Alignment (MSA)

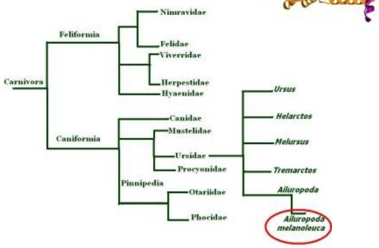
INTRODUCTION


Evolutionary
QFGLFSPEEIRASSVALIR--YPETI NG--VPKESGLVCAGHFGHIELVK
QFGLFSPEEIKRMSVVHVE--YPET QRQRPRTKGLECPGHFGHIELAT
ELGVLDPEIIKKISVCEIV--NV --FPREGGLYCPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKY GRPKLGGLECPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIK RPKLGGLECPGHFGHIELAK
QFGILSPDEIRRMSVTEGGV KLGGLECPGHFGHIDLAK
QFGILGPEEIKRMSVAH-- LGGLDCPGHFGHLELAK
QFGILSPEEIRSMSVAK- GGLDCPGHFGHIELAK
QFGILSPDEIRQMSVIH GLECPGHFGYLELAK
---------------- -ECPGHFGHIELAK
-------------- ECPGHFGFIELAK
QFEIFKERQIKSY CPGHFGYIELAE
QFEVFKEAQIKA GHFGYVELAE
MSA
Structural
Functional

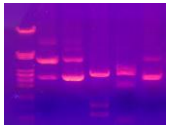
Motif and Domain Identification



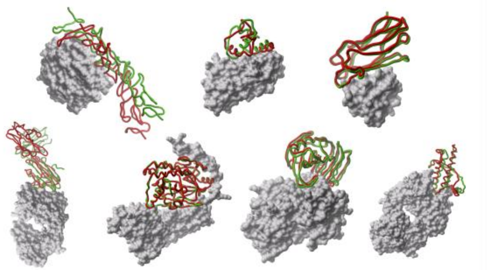
Phylogenetics



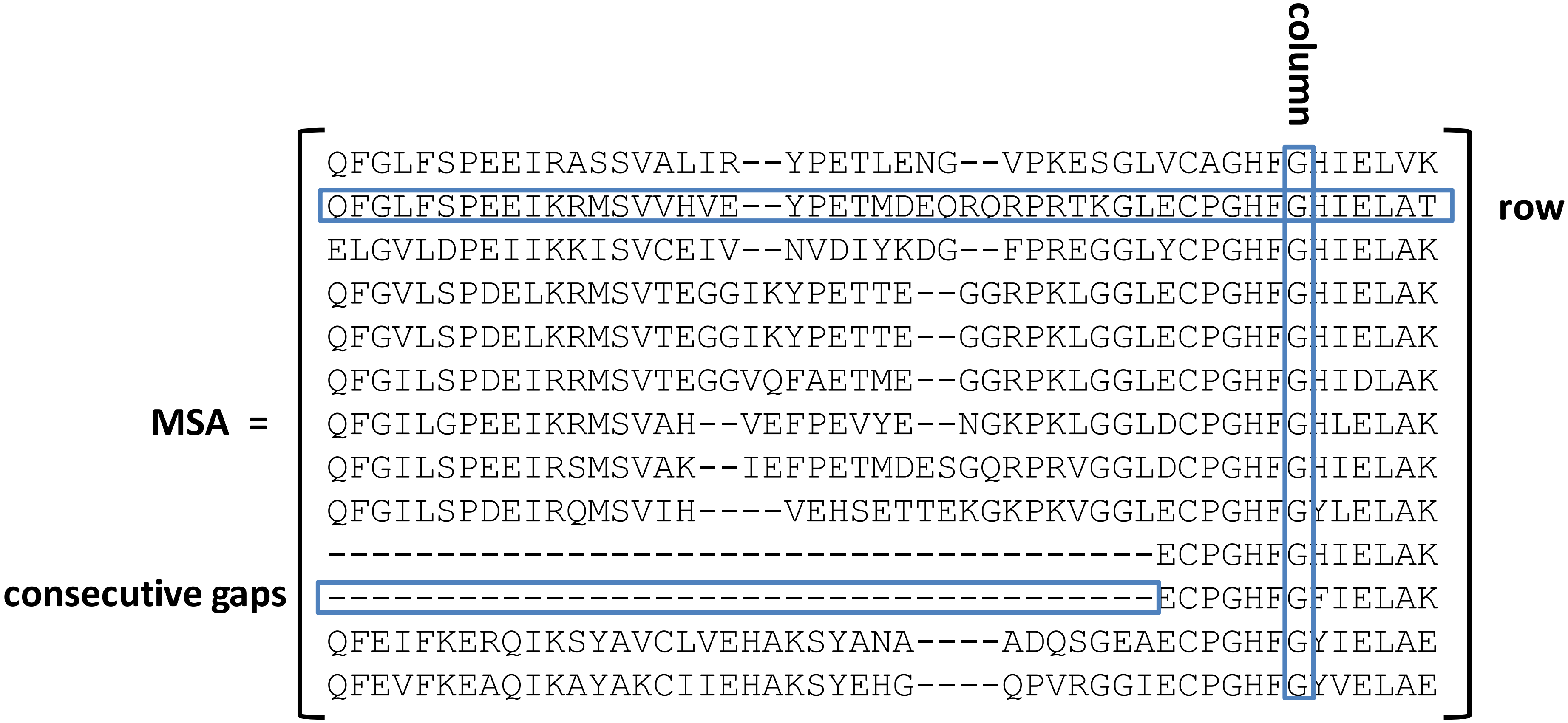
Physiological Studies



Protein Structure and Interaction



INTRODUCTION



column

row

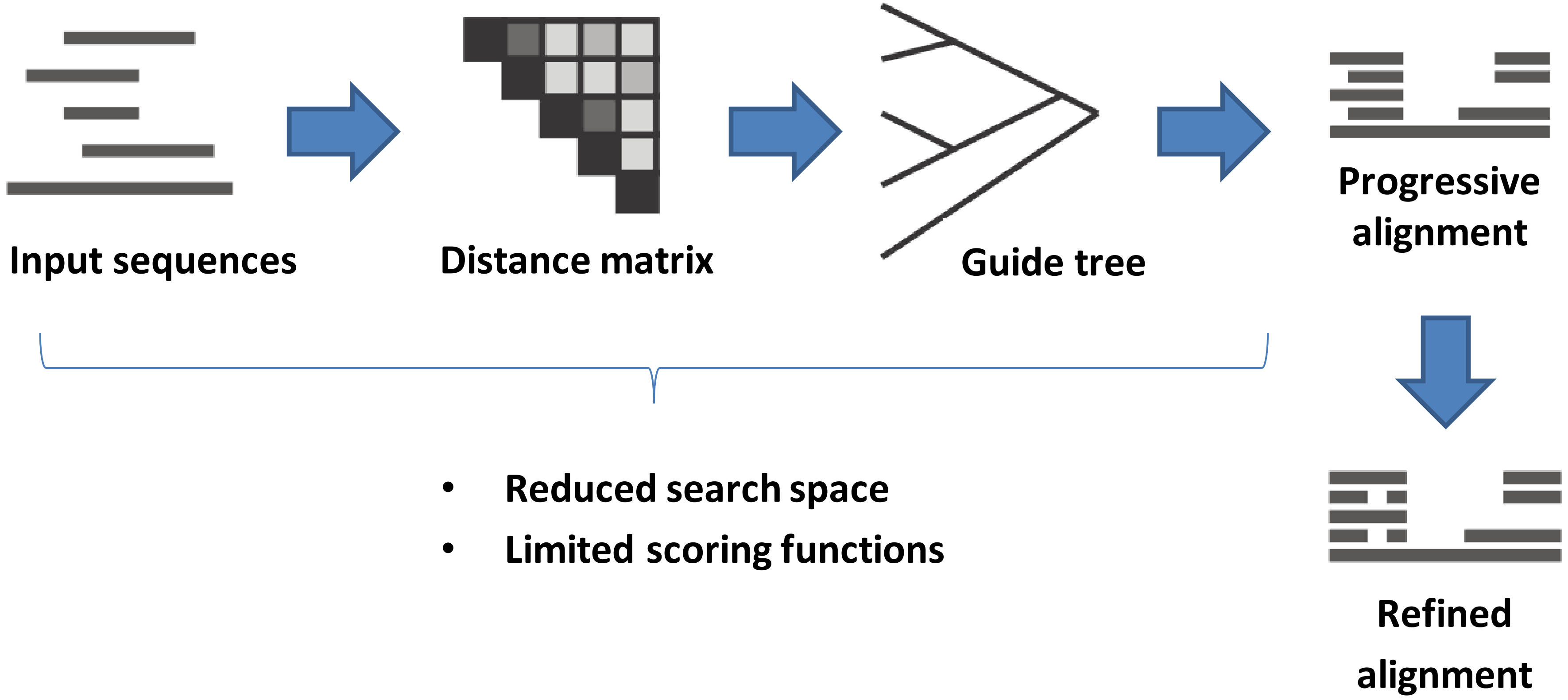
MSA =

consecutive gaps

```
QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAGHFGHIELVK
QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
QFGILSPDEIRQMSVIH----VEHSETTEKGKPKVGGLECPGHFGYLELAK
--------------------------------------ECPGHFGHIELAK
--------------------------------------ECPGHFGFIELAK
QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
QFEVFKEAQIKAYAKCIIEHAKSYEHG----QPVRGGIECPGHFGYVELAE
```

INTRODUCTION

MSA algorithms workflow:



Input sequences

Distance matrix

Guide tree

Progressive alignment

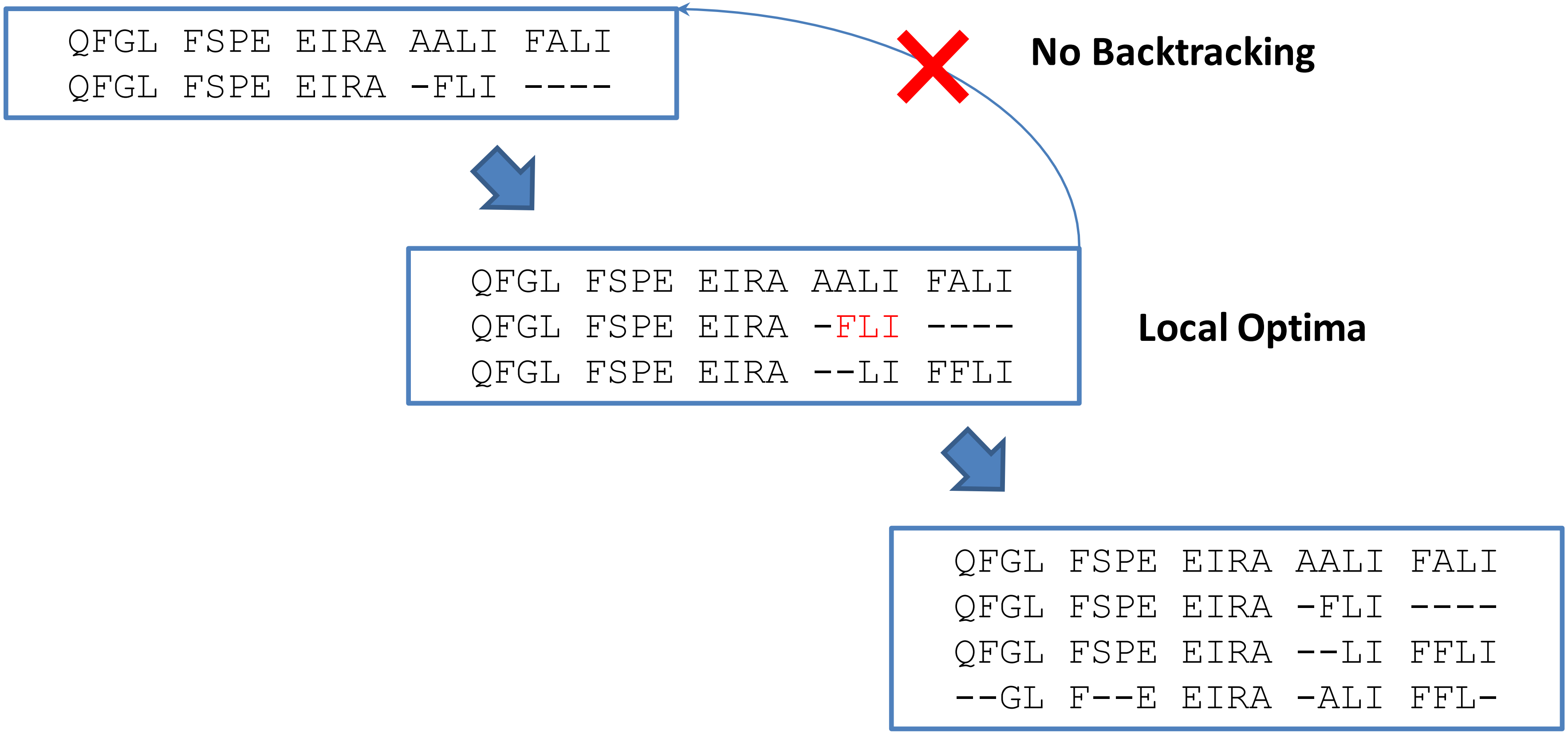
Refined alignment

- Reduced search space
- Limited scoring functions

Do, C. B., Katoh, K. 2008. Protein Multiple Sequence Alignment. *Functional Proteomics: Methods and Protocols*, *484*, 379-413.

INTRODUCTION

Progressive *vs* Iterative Methods:



```
QFGL FSPE EIRA AALI FALI
QFGL FSPE EIRA -FLI ----
```

No Backtracking

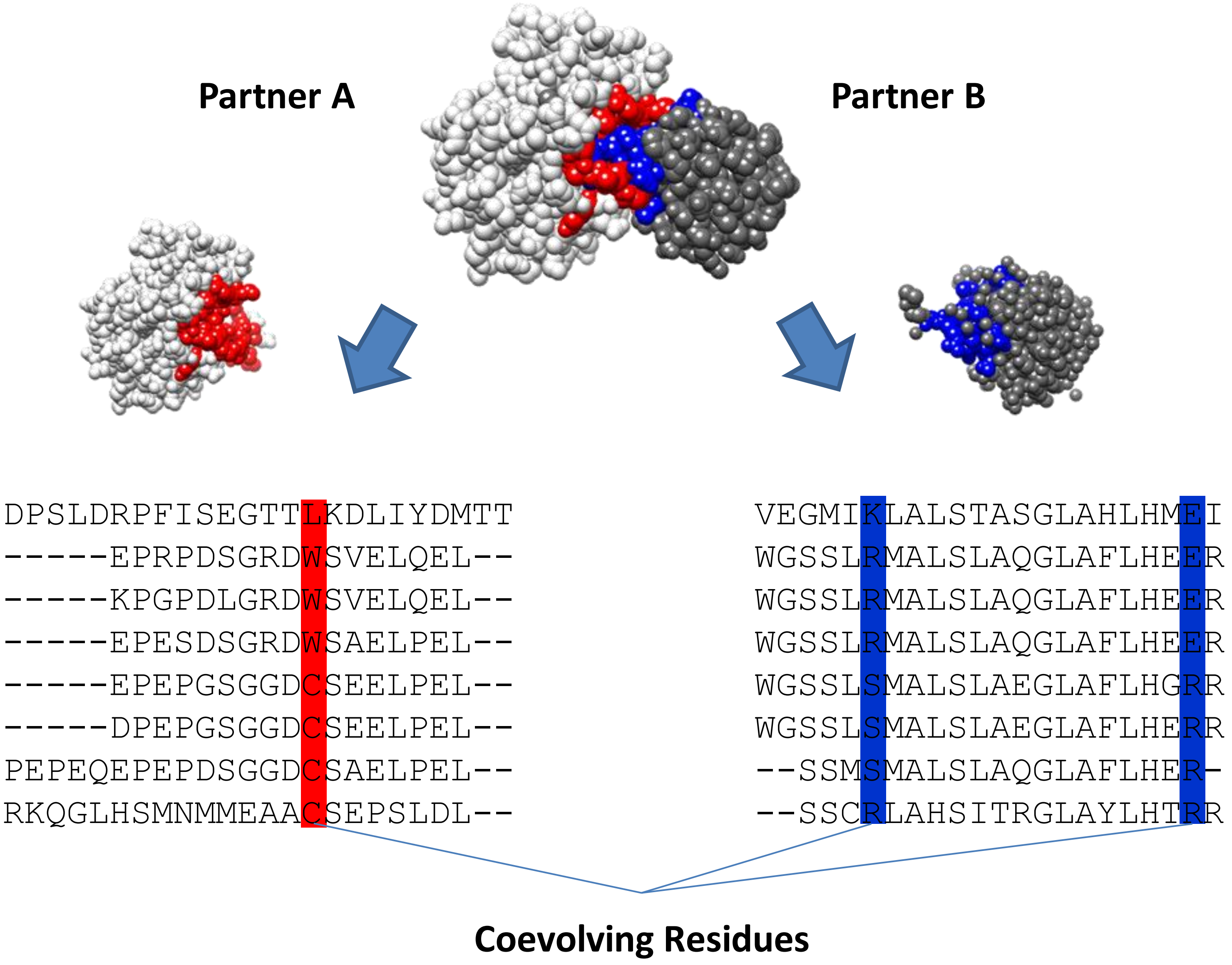
```
QFGL FSPE EIRA AALI FALI
QFGL FSPE EIRA -FLI ----
QFGL FSPE EIRA --LI FFLI
```

Local Optima

```
QFGL FSPE EIRA AALI FALI
QFGL FSPE EIRA -FLI ----
QFGL FSPE EIRA --LI FFLI
--GL F--E EIRA -ALI FFL-
```

Notredame, C. 2002. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, *3*(1), 131-44.

INTRODUCTION



```
DPSLDRPFISEGTTLKDLIYDMTT
-----EPRPDSGRDWSVELQEL--
-----KPGPDLGRDWSVELQEL--
-----EPESDSGRDWSAELPEL--
-----EPEPGSGGDCSEELPEL--
-----DPEPGSGGDCSEELPEL--
PEPEQEPEPDSGGDCSAELPEL--
RKQGLHSMNMMEAACSEPSLDL--
```

```
VEGMIKLALSTASGLAHLHMEI
WGSSLRMALSLAQGLAFLHEER
WGSSLRMALSLAQGLAFLHEER
WGSSLRMALSLAQGLAFLHEER
WGSSLSMALSLAEGLAFLHGRR
WGSSLSMALSLAEGLAFLHERR
--SSMSMALSLAQGLAFLHER-
--SSCRLAHSITRGLAYLHTRR
```

Coevolving Residues

INTRODUCTION

```
QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAGHFGHIELVK
QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
QFGILSPDEIRQMSVIH----VEHSETTEKGKPKVGGLECPGHFGYLELAK
--------------------------------------ECPGHFGHIELAK
--------------------------------------ECPGHFGFIELAK
QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
QFEVFKEAQIKAYAKCIIEHAKSYEHG----QPVRGGIECPGHFGYVELAE
```

Manual fixing of misalignments:

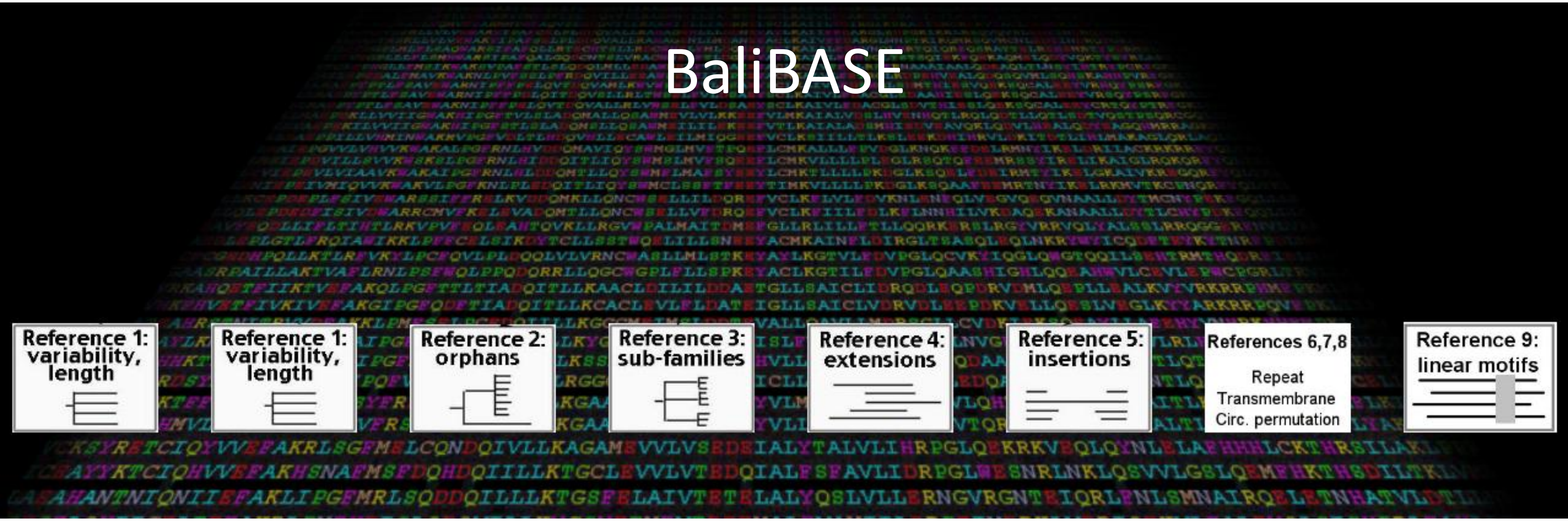
Evolutionary sense: Only two insertions

```
QFGLFSPEEIRASSVAL--IRYPETLE--NGVPKESGLVCAGHFGHIELVK
QFGLFSPEEIKRMSVVH--VEYPETMDEQRQRPRTKGLECPGHFGHIELAT
QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
ELGVLDPEIIKKISVCE--IVNVDIYK--DGFPREGGLYCPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
QFGILSPDEIRQMSVIH--VEHSETTE--KGKPKVGGLECPGHFGYLELAK
--------------------------------------ECPGHFGHIELAK
--------------------------------------ECPGHFGFIELAK
QFEIFKERQIKSYAVCL--VEHAKSYA--NAADQSGEAECPGHFGYIELAE
QFEVFKEAQIKAYAKCI--IEHAKSY--EHGQPVRGGIECPGHFGYVELAE
```

Baldauf, S. 2003. Phylogeny for the faint of heart: a tutorial. *Trends in Genetics*, *19*(6), 345-351.

INTRODUCTION

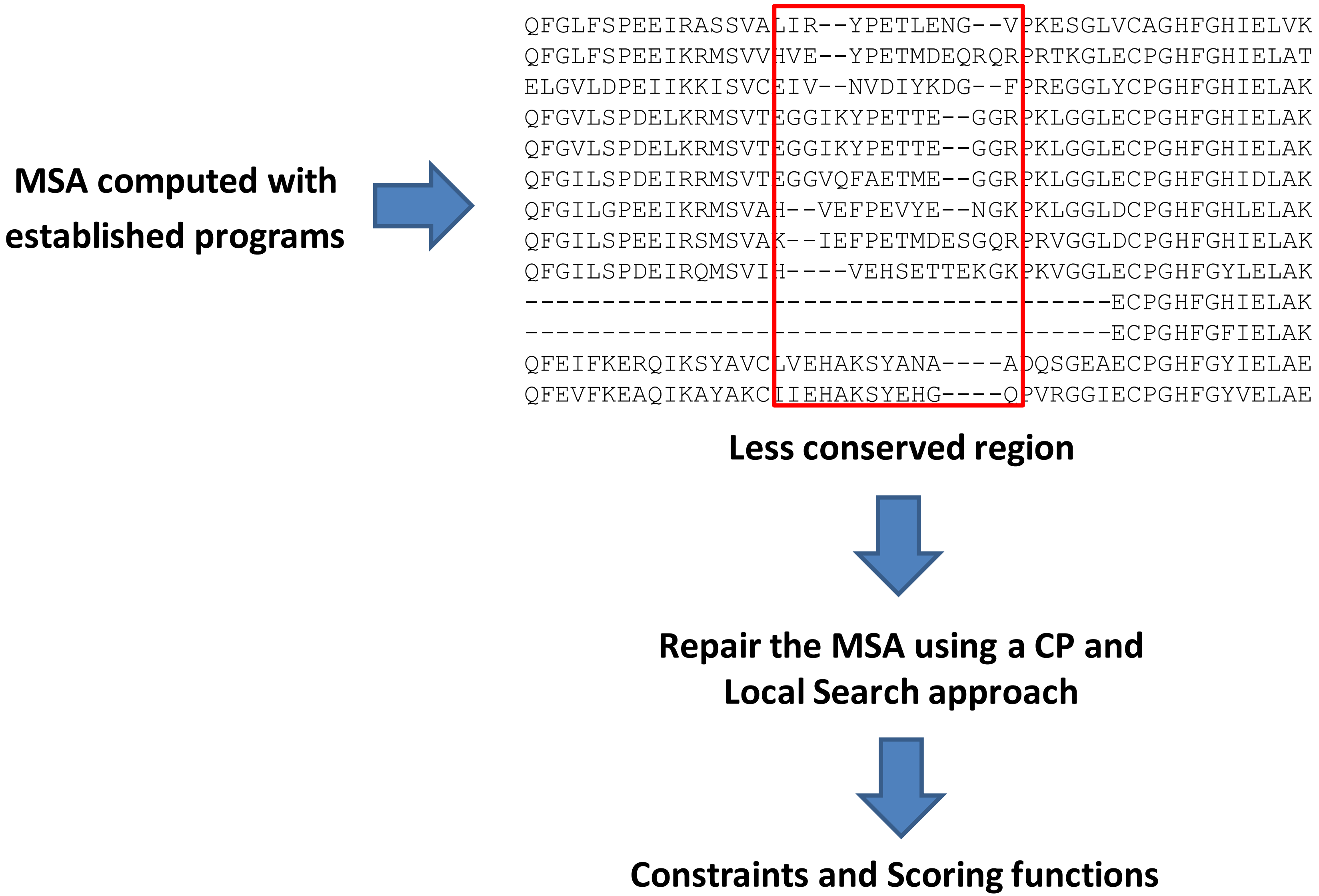
Evaluation and assessment of MSA:



BAliBASE provides high quality alignments, manually refined and based on 3D structural superpositions

Thompson, J. D., Koehl, P., Ripp, R., Poch, O. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, *61*(1), 127-36. doi:10.1002/prot.20527

OUR APPROACH



MSA computed with established programs

```
QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAGHFGHIELVK
QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
QFGILSPDEIRQMSVIH----VEHSETTEKGKPKVGGLECPGHFGYLELAK
--------------------------------------ECPGHFGHIELAK
--------------------------------------ECPGHFGFIELAK
QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
QFEVFKEAQIKAYAKCIIEHAKSYEHG----QPVRGGIECPGHFGYVELAE
```

Less conserved region

Repair the MSA using a CP and Local Search approach

Constraints and Scoring functions

OUTLINE

1. Introduction
   - i. Our proposal
2. Method
   - i. CP Model
   - ii. Search
   - iii. Local Search
   - iv. Experiments
3. Results and Discussion
4. Conclusions

METHOD

p

n

S	P	V	I	-	-	-	L
R	-	-	I	-	-	-	S
S	-	-	-	-	-	-	L
F	N	T	T	Q	G	G	P
T	-	-	-	-	-	-	-
F	S	K	N	-	-	-	-
E	T	F	G	Q	-	-	-
K	S	-	-	-	-	-	T

$a_{i,j}$ - residue at sequence ***i*** and position ***j***

$\mathbf{s}_i = \langle a_{i,1}, \ldots, a_{i,p} \rangle$ - sequence ***i***

$\sigma_A \left(a_{1,j}, a_{2,j} \right)$ - scoring function

g_i - number of gaps in $\mathbf{s}_i$

$\gamma \left(\mathbf{s}_1, \mathbf{s}_2 \right)$ - gap penalty

METHOD

Score for the alignment of two sequences, s_1 and s_2:

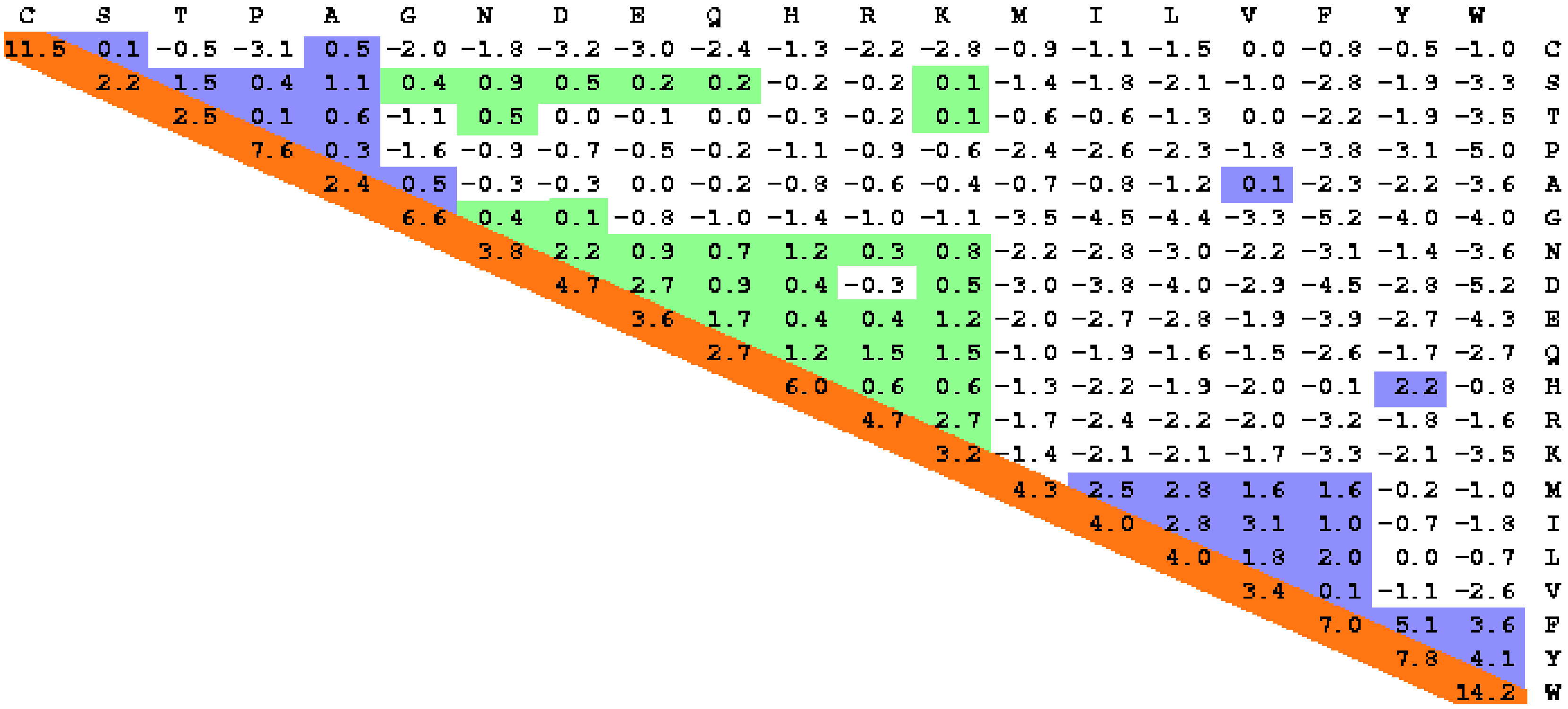
$$\sigma_S\left(\mathbf{s}_1, \mathbf{s}_2\right) = \sum_{i=1}^{p} \underbrace{\sigma_A\left(a_{1,i}, a_{2,i}\right)}_{\textbf{Substitution Matrix}} - \gamma\left(\mathbf{s}_1, \mathbf{s}_2\right)$$

The Score for the multiple alignment is derived from pairwise scores:

$$\sigma = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sigma_S\left(\mathbf{s}_i, \mathbf{s}_j\right)$$

METHOD

GONNET substitution matrix:



C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
11.5	0.1	-0.5	-3.1	0.5	-2.0	-1.8	-3.2	-3.0	-2.4	-1.3	-2.2	-2.8	-0.9	-1.1	-1.5	0.0	-0.8	-0.5	-1.0	C
	2.2	1.5	0.4	1.1	0.4	0.9	0.5	0.2	0.2	-0.2	-0.2	0.1	-1.4	-1.8	-2.1	-1.0	-2.8	-1.9	-3.3	S
		2.5	0.1	0.6	-1.1	0.5	0.0	-0.1	0.0	-0.3	-0.2	0.1	-0.6	-0.6	-1.3	0.0	-2.2	-1.9	-3.5	T
			7.6	0.3	-1.6	-0.9	-0.7	-0.5	-0.2	-1.1	-0.9	-0.6	-2.4	-2.6	-2.3	-1.8	-3.8	-3.1	-5.0	P
				2.4	0.5	-0.3	-0.3	0.0	-0.2	-0.8	-0.6	-0.4	-0.7	-0.8	-1.2	0.1	-2.3	-2.2	-3.6	A
					6.6	0.4	0.1	-0.8	-1.0	-1.4	-1.0	-1.1	-3.5	-4.5	-4.4	-3.3	-5.2	-4.0	-4.0	G
						3.8	2.2	0.9	0.7	1.2	0.3	0.8	-2.2	-2.8	-3.0	-2.2	-3.1	-1.4	-3.6	N
							4.7	2.7	0.9	0.4	-0.3	0.5	-3.0	-3.8	-4.0	-2.9	-4.5	-2.8	-5.2	D
								3.6	1.7	0.4	0.4	1.2	-2.0	-2.7	-2.8	-1.9	-3.9	-2.7	-4.3	E
									2.7	1.2	1.5	1.5	-1.0	-1.9	-1.6	-1.5	-2.6	-1.7	-2.7	Q
										6.0	0.6	0.6	-1.3	-2.2	-1.9	-2.0	-0.1	2.2	-0.8	H
											4.7	2.7	-1.7	-2.4	-2.2	-2.0	-3.2	-1.8	-1.6	R
												3.2	-1.4	-2.1	-2.1	-1.7	-3.3	-2.1	-3.5	K
													4.3	2.5	2.8	1.6	1.6	-0.2	-1.0	M
														4.0	2.8	3.1	1.0	-0.7	-1.8	I
															4.0	1.8	2.0	0.0	-0.7	L
																3.4	0.1	-1.1	-2.6	V
																	7.0	5.1	3.6	F
																		7.8	4.1	Y
																			14.2	W

CP MODEL

Variables:

$x_{i,j} \in X$ - for each cell in the matrix with domain

$$D\left(x_{i,j}\right) = \{'-',' A',' C',' D', \ldots ,' Y',' W'\}$$

$x_{i,k}^{G} \in X_i^G$ - models the position of the *k*'th gap in *S_i*; $1 \cdot\ k \cdot\ g_i$

c_i - specify the number of consecutive gaps

Constraint:

$$\textsc{validSequence}(\mathbf{s}_i) = \textsc{inTable}\left(\langle x_{i,1}, \ldots , x_{i,p}, c_i \rangle , T_i\right)$$

each sequence is obtained by changing the positions of the gaps

each table is created so that each row is obtained by placing the gaps in distinct positions

$$T_i = C_{g_i}^{p}$$

SEARCH

Use of a greedy variable and value heuristics:

$$
\begin{aligned}
var\left(X\right) &= \arg\max_{x_{i,j}} \max_{v_1, v_2 \in D(x_{i,j})} q\left(x_{i,j}, v_1\right) - q\left(x_{i,j}, v_2\right) \\
val\left(x_{i,j}\right) &= \arg\max_{v \in D(x_{i,j})} \underbrace{q\left(x_{i,j}, v\right)}
\end{aligned}
$$

$$q\left(x_{i,j}, v\right) = q^A\left(x_{i,j}, v\right) + nq^G\left(x_{i,j}, v\right)$$

- function which estimates the cost of assigning value *v* to variable $x_{i,j}$

SEARCH

$$q\left(x_{i,j},v\right) = \underbrace{q^{A}\left(x_{i,j},v\right)}_{\text{(A)}} + \underbrace{nq^{G}\left(x_{i,j},v\right)}_{\text{(B)}}$$

- function which estimates the cost of assigning value *v* to variable $x_{i,j}$

(A) Cost based on the set of residues already assigned:

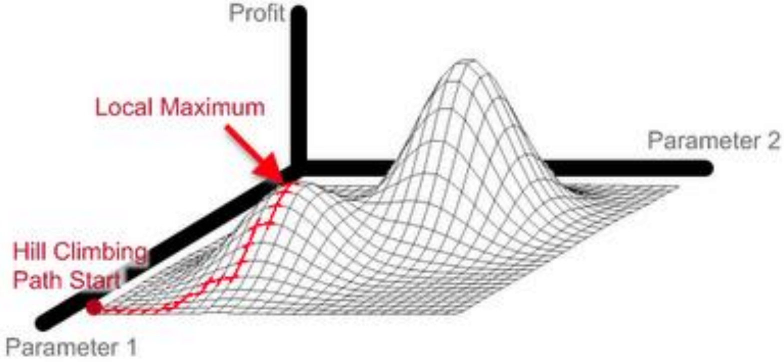
$$q^{A}\left(x_{i,j},v\right) = \sum_{k=1}^{n} \begin{cases} \sigma_{A}\left(x_{k,j},v\right) & \Leftarrow \left|D\left(x_{k,j}\right)\right| = 1 \\ 0 & \Leftarrow \text{otherwise} \end{cases}$$

(B) Cost based on the number of consecutive gaps:

$$nq^{G}\left(x_{i,j},v\right) = \begin{cases} -10 & \Leftarrow \text{if it creates a new gap} \\ 10 & \Leftarrow \text{if it does not creates a new gap} \\ 0 & \Leftarrow \text{not known} \end{cases}$$

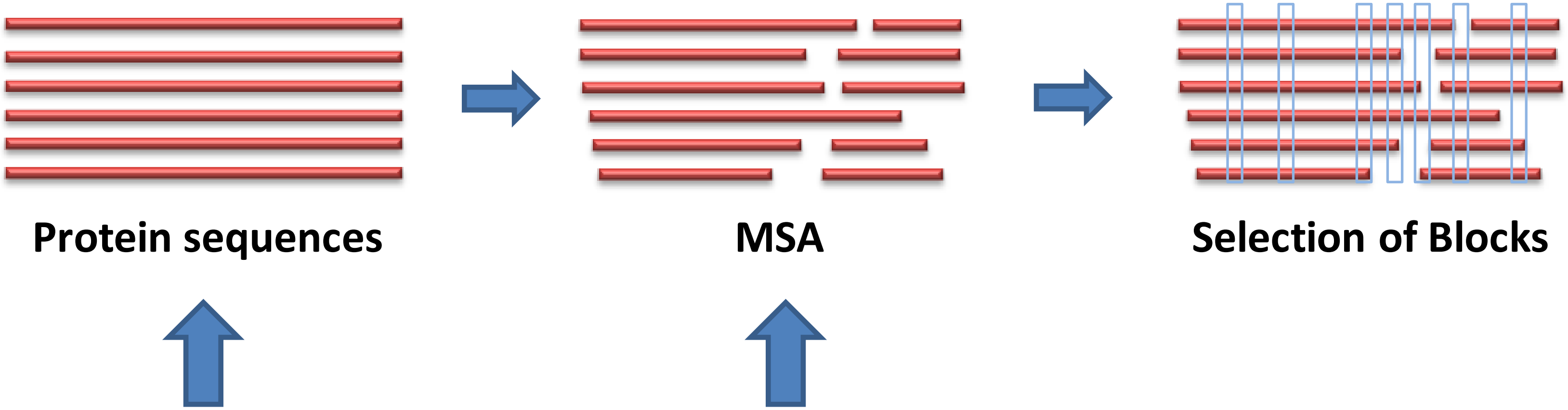
LOCAL SEARCH

The same model was tested using a greedy hill-climbing heuristic optimization:



- **Constrained local search (COMET)**
- **The same objective function**
- **Randomization of start point**

EXPERIMENTS



Protein sequences

MSA

Selection of Blocks

Selected examples from BaliBASE

ClustalW with GONNET matrix

EXPERIMENTS

Example of a Block:

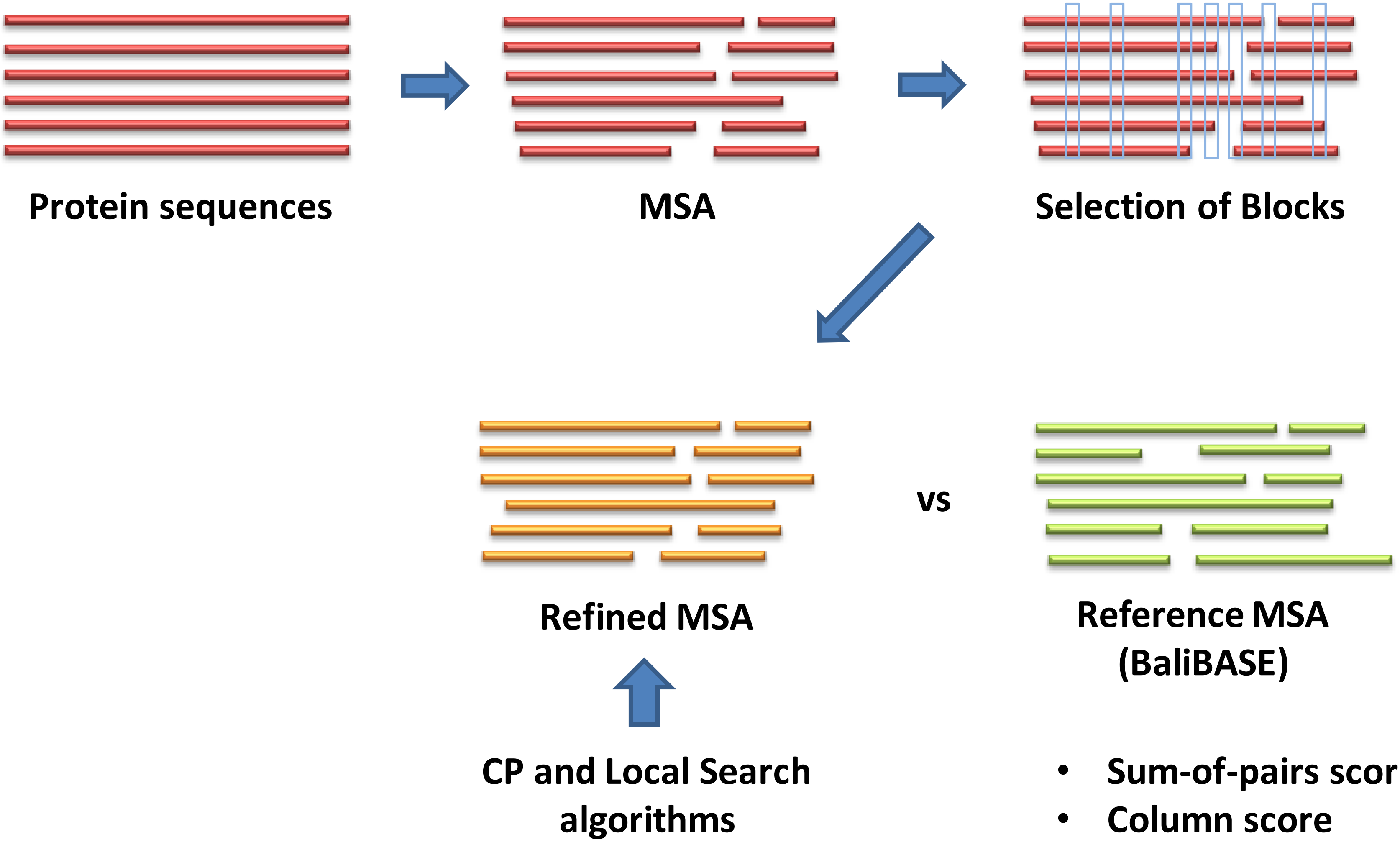
```
QNLDFAVALPAINVAAFSKN--------TTRKLELAVQNMSQFEKGAYTGEISAQ
DNLDFAIAPSFTSLALISTS-------KID-KLKVAAQNLSQFDSGAFTGEISAK
AETEALVCVPATLLSHAAE-----IL-RT--PVHAGGEDCHTKESGAYTGCISAE
ARVDALICPPATLLYVATA-----LC-DS--PLMIGAQDCHQKQSGAHTGEVSAE
RLFEALICVPATLLSRAFD-----IL-GE--NILLGGQNCHFDDYGPYTGDISAF
PQIITGIIPPFTLLSAQQAV------SDS--PIFLGAQTTHEADSGAFTGEISAP
PKIITGIIPPFTLLSSQQII------KNT--PIRLGAQTLHEVDSGAFTGEISAP
LSCTIGIASPFTSLRAIHEMI-----NTTG-FLWLGAQNVHPELSGAFTGEISLP
KAVLGIAPVHVHLTEVNKVLP--------N-NLLLLAQDANFIASGSYTGTVSYT
VNADYSVGVPSIYLNQAKEI-------LKG--IKVIAQDAHFKNEGAYTGNISWS
RVLIGLAAPTVYLLQLHN---AMQIVLN-NRILTCAQDVSRFPDNGAYTGEVSAE
EKNTVIIAPPTIYLERVCKNIS-------NMNIFLGSQNVDINLNGAFTGETSIL
KNNIIIIAPPTVFLERVYKDIN--------INIHLAAQNIDVNLTGAFTGENSAL
```

EXPERIMENTS

CP and Local Search algorithms:

```
SKN--------TTRKL      SKN--------TTRKL
STS-------KID-KL      STS-------KID-KL
AE-----IL-RT--PV      AE-I----L-RT--PV
TA-----LC-DS--PL      TAL-----LCDS--PL
FD-----IL-GE--NI      FDII----L-GE--NI
QQAV-----SDS--PI      QQAV-----SDS--PI
QQII-----KNT--PI  →   QQII-----KNT--PI
HEMI-----NTTG-FL      HEMI------NTTGFL
KVLP--------N-NL      KVLP--------N-NL
KEI-------LKG--I      KE-I------LKG--I
N---AMQIVLN-NRIL      NA--MQIVL-N-NRIL
CKNIS-------NMNI      CKNI-------SNMNI
YKDIN--------INI      YKDI--------NINI
```

EXPERIMENTS



Protein sequences

MSA

Selection of Blocks

Refined MSA

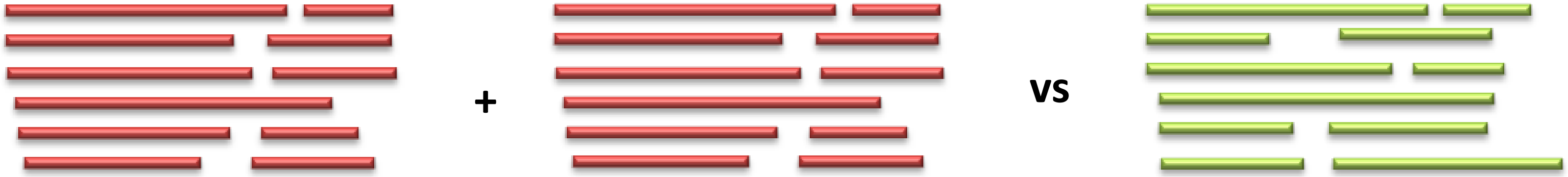
vs

Reference MSA (BaliBASE)

CP and Local Search algorithms

- Sum-of-pairs score
- Column score

EXPERIMENTS



+ vs

CP :
22 alignments

Local Search:
24 alignments

Reference MSA
(BaliBASE)

Sum-of pairs score (SP):

The percentage of correctly aligned pairs of residues in the test alignment, relative to the reference alignment.

Column score (CS):

The percentage of correctly aligned columns, which tests the ability of the programs to align all of the sequences correctly at any given position.

OUTLINE

1. Introduction
    i. Our proposal
2. Method
    i. CP Model
    ii. Search
    iii. Local Search
    iv. Experiments
3. Results and Discussion
4. Conclusions

RESULTS AND DISCUSSION

CP:

- Average improvement per column above the average score attributed to a match
- Average number of columns was 54 (in 22 MSA)
- 77% of MSAs (17 out of 22) were improved
- Average improvement of 11% in the SP score
- Several minutes per alignment

Local Search:

- Insignificant average improvement
- 58% of MSAs (14 out of 24) were improved
- Few seconds per alignment

OUTLINE

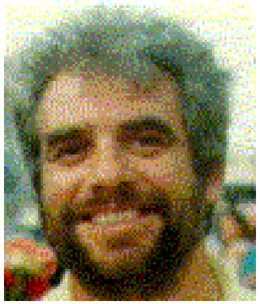
1. Introduction
   i. Our proposal
2. Method
   i. CP Model
   ii. Search
   iii. Local Search
   iv. Experiments
3. Results and Discussion
4. Conclusions

CONCLUSIONS

A framework for improving MSA obtained with classic algorithms

- **Allows the use of different scoring functions**
- **Too much room for optimizations, both on the heuristics and on the scoring functions**
- **Possible inclusion of structural information**

ACKNOWLEDGEMENTS



Pedro Barahona, PhD

Ludwig Krippahl, PhD

Marco Correia, PhD

This work was funded by FCT, under project PTDC/EIA-CCO/115999/2009.

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR