

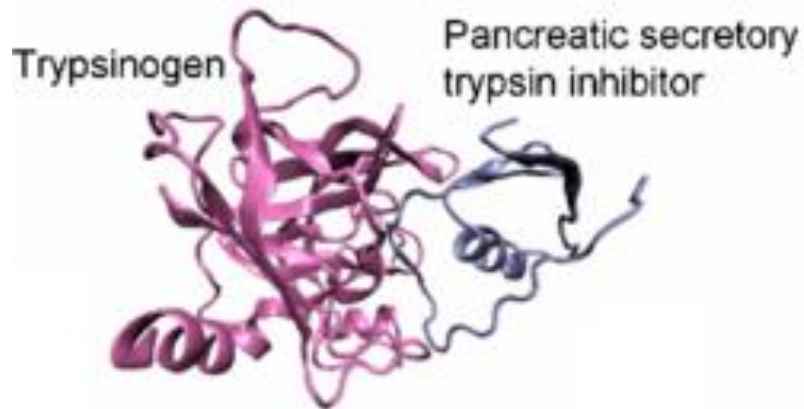
Multiple Sequence Alignment

- a quick overview -

Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLDDYFKCFIVGADNVGSKOMQIIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLDDYFKCFIVGADNVGSKOMQIIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLDDYFKCFIVGADNVGSKOMQIIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLDDYFKCFIVGADNVGSKOMQIIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFKIIQLDDYFKCFVVGADNVGSKOMQIIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKSNYFLKIIQLDDYFKCFIVGADNVGSKOMQIIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
Q7ZUG3_BRARE	-----MPREDRATWKSNYFLKIIQLDDYFKCFIVGADNVGSKOMQIIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0 ICTPU	-----MPREDRATWKSNYFLKIIQLDDYFKCFIVGADNVGSKOMQIIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0 DROME	-----MVRENKAANKAQYFIKVVLFDEFKCFIVGADNVGSKOMQIIRMSLRGL-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0 DICDI	-----MSGAG-SKRKKLFIEKATKLFITTDKMIVAEADFYGSOLQKIRKSIIRGI-GAYLMGKNTMIRKIVIRDLADSK--PELD	75
Q54LP0_DICDI	-----MSGAG-SKRKNVFIKATKLFITTDKMIVAEADFYGSOLQKIRKSIIRGI-GAYLMGKNTMIRKIVIRDLADSK--PELD	75
RLA0 FLAFB	-----MAKLSKQKKQMYIEKLSLIQQYSKILVHVQNVGSKOMASVKKSLRGK-ATILMGKNTIRITALKKNLAV--DQIE	76
RLA0 SULAC	-----MIGLAYTTTKKIANKVYDEVAELTEKLTHTKTIITANIEGFPADKLHEIRKKLRGK-ADIKVTKNLFPNIALKNAG----DYIK	79
RLA0 SULTO	-----MRIMAVITQERKIANKKIEEVKELEKLRHYHTIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNLTFGIAAKNAG----LDVS	80
RLA0 SULSO	-----MKRLALALKQRKVASWKEEVKELTELIKMSNTILIGNIEGFPADKLHEIRKKLRGK-ATIKVTKNLTFKIAAKNAG----IDIE	80
RLA0 AERPE	MSVYSIVGQMYKREKPIPEMKTLMLELEELFSKRVVLFADLTGTPFVVDVVRKKLWKK-YPMVAKKRIILHAMKAAGLE--LDDN	86
RLA0 PYRAE	-----MHLAIGKRRYVTRQYPAKRYKIYSEATELLQKIPYVFLFDLHGLSIRILHEVRYRLRRY-GVIRIKKPIFLFKIAFTKVYGG--IPAE	85
RLA0 METAC	-----MAEERNHTEHIPQWKDEIENIKELIQSHKVFQGVGIEGILATKMDKIRRDLDKV-AVLKVRNTLTLEHALNQLG--ETIP	78
RLA0 METMA	-----MAEERNHTEHIPQWKDEIENIKELIQSHKVFQGVGIEGILATKMDKIRRDLDKV-AVLKVRNTLTLEHALNQLG--ESIP	78
RLA0 ARCFU	-----MAAVRGS--PPEYKVRAVEEIKRMISSEKPYVAIVSFRNVFAGOMQIRREFRGK-AEIKVKNLTLEHALDALG--GDYL	75
RLA0 METKA	MAVKAKGQPPFSQYEIPKVAEWKRREVKELELMDEYENYGLVDLEGIPAPQLQEIIRAKLRERDIIIRMRNTLHRIALEEKLDER--PELE	88
RLA0 METTH	-----MAHVAEWKKKEVQELNDLIKSEYVYCIANLADIPAROLQKMRQTLRDS-ALIRMRKKTLLISLAEKAGREL--EHVD	74
RLA0 METTL	-----MITAESENKIAPWKIEEVNKLKELLKNGQIYALVDMMEVPAROLQEIIRDKIR-ETMTLKMRNTLTLEHAKEVAEETGNDEFA	82
RLA0 METVA	-----MIDAKSENKIAPWKIEEVNALKELLKSANVIALIDHMEVPAROLQEIIRDKIR-DQMTLKMRNTLTLEHAKEVAEETGNDEFA	82
RLA0 METJA	-----METKVKANVAPWKIEEVKTLKGLIKSKPYVAIVDMDVDPAPQLQEIIRDKIR-DKVKLRMRNTLTLEHALKEAAEELNNPKLA	81
RLA0 PYRBA	-----MAHVAEWKKKEVEELANLIKSPYVIALVDVSSMPAYPLSQMRRLIRENGGLLRVRNTLTLELAIKKAAEELGKPELE	77
RLA0 PYRHO	-----MAHVAEWKKKEVEELAKLIKSPYVIALVDVSSMPAYPLSQMRRLIRENGGLLRVRNTLTLELAIKKAAEELGKPELE	77
RLA0 PYRFU	-----MAHVAEWKKKEVEELANLIKSPYVIALVDVSSMPAYPLSQMRRLIRENGGLLRVRNTLTLELAIKKAAEELGKPELE	77
RLA0 PYRKO	-----MAHVAEWKKKEVEELANLIKSPYVIALVDVAGVPAYPLSKMRDELE-CKALLRVRNTLTLELAIKRAAEELGQPELE	76
RLA0 HALMA	MSAESERKTETIPEWKQEEVDIYFMIESYESYGVVNIAGIPHROLQDMRRDLHGT-AELRVVRNTLTLEHALDDVD--DGLE	79
RLA0 HALVO	MSSEVVRQTEVIPQWKREEVDYDFIESYESYGVVNVGAGIPHROLQDMRRDLHGT-AAVRMRNTLTLYNHALDEVN--DGFE	79
RLA0 HALSA	MSAEERQRTTEVPENKRQEVAEVLDLLETDSYGVVNVGAGIPHROLQDMRRDLHGT-AALRMRNTLTLYNHALEEA--DGLD	79
RLA0 THEAC	-----MKEVSOQKKELVNEITRIKASRSVAIVDTAGIRHROIDIRGKNRGK-INLEVIKKTLLFKALENLGD--EKLK	72
RLA0 THEVO	-----MRKINPKKEIVSELAADITKSKAVAIVDIKCYRROMODIRAKNRDK-VKIKVVKKILLFKALDSIND--EKLK	72
RLA0 PICTO	-----MTEPAQWKIDFVKKLENEINSRKVAIVSISKGLRNNTFOKIRNSIRDK-ARIKVRARLLRLAIENTGK--NHIV	72
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90	

Motivations for sequence alignment

1) find genes that are related by common descent



Chymotrypsin	VKKTMCAGG-DGVISACNGDSGGPLNCQLENGSWEVFGIVSFGSRRGC [...]
	+ M C G +G +C GDSGGP+ C NG + G+VS+G GC
Trypsinogen	ITSNMFCVGFLEGGKDSCQGDGGPVVC---NGQLQ--GVVSWGD--GC [...]

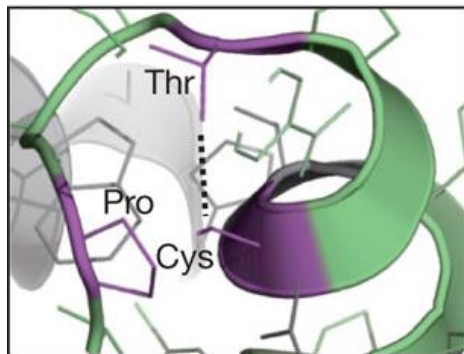
Motivations for sequence alignment

2) to identify and check the state of “active sites”

Ot	Q	D	I	K	L	S	D	Y	R	G	-	-	K	Y	V	V	L	F	F	Y	P	L	D	F	T	F	V	C	P	T	E	I	T	A	F	S	D	R	Y	E	E	F	A	K	L	N	T	E	V	L	G	V	S	V
Se	Q	T	I	K	L	S	N	Y	R	G	-	-	K	Y	V	V	L	F	F	Y	P	L	D	F	T	F	V	C	P	T	E	I	T	A	F	S	D	R	Y	A	D	F	S	A	L	N	T	E	I	L	G	V	S	V
At	I	K	V	K	L	S	D	Y	N	G	K	-	K	Y	V	I	L	F	F	Y	P	L	D	F	T	F	V	C	P	T	E	I	T	A	F	S	D	R	H	S	E	F	E	K	L	N	T	E	V	L	G	V	S	V
Hs	K	E	V	K	L	S	D	Y	K	G	-	-	K	Y	V	V	L	F	F	Y	P	L	D	F	T	F	V	C	P	T	E	I	I	A	F	S	N	R	A	E	D	F	R	K	L	G	C	E	V	L	G	V	S	V
Mm	K	E	I	K	L	S	D	Y	R	G	-	-	K	Y	V	V	L	F	F	Y	P	L	D	F	T	F	V	C	P	T	E	I	I	A	F	S	D	H	A	E	D	F	R	K	L	G	C	E	V	L	G	V	S	V
Ce	V	D	V	S	L	S	D	Y	K	G	-	-	K	Y	V	V	L	F	F	Y	P	L	D	F	T	F	V	C	P	T	E	I	I	A	F	S	D	R	A	E	E	F	K	A	I	N	T	V	V	L	A	A	S	T
Se	D	E	V	S	L	D	K	Y	K	G	-	-	K	Y	V	V	L	A	F	I	P	L	A	F	T	F	V	C	P	T	E	I	I	A	F	S	E	A	A	K	K	F	E	E	Q	G	A	Q	V	L	F	A	S	T
Dm	K	D	I	K	L	S	D	Y	K	G	-	-	K	Y	L	V	L	F	F	Y	P	L	D	F	T	F	V	C	P	T	E	I	I	A	F	S	E	S	A	A	E	F	R	K	I	N	C	E	V	I	G	C	S	T
Nc	-	P	I	D	F	H	E	F	I	G	D	-	N	W	V	I	L	F	S	H	P	E	D	Y	T	P	V	C	T	T	E	L	G	E	M	A	R	L	E	P	E	F	K	K	R	G	V	K	L	I	G	L	S	A
Has	T	R	L	G	L	T	D	A	L	A	D	N	R	A	V	V	L	F	F	Y	P	F	D	F	S	P	V	C	A	T	E	L	C	A	I	Q	N	A	R	W	F	D	C	T	P	G	L	A	V	W	G	I	S	P

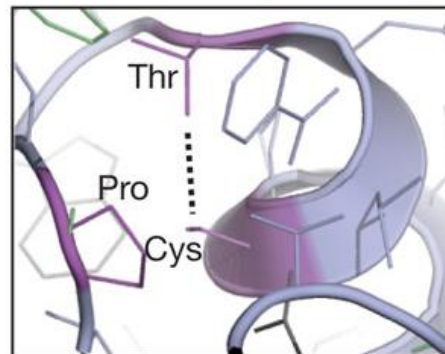
b

Generic
(PRX-V)



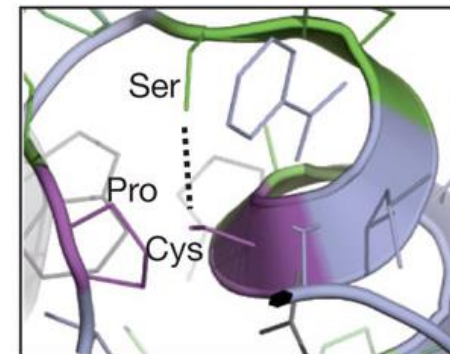
PGAFTPG**C**SKTH

Human
(PRDX2)



PLDFTFV**C**PTEI

Archaea
(HyrA)

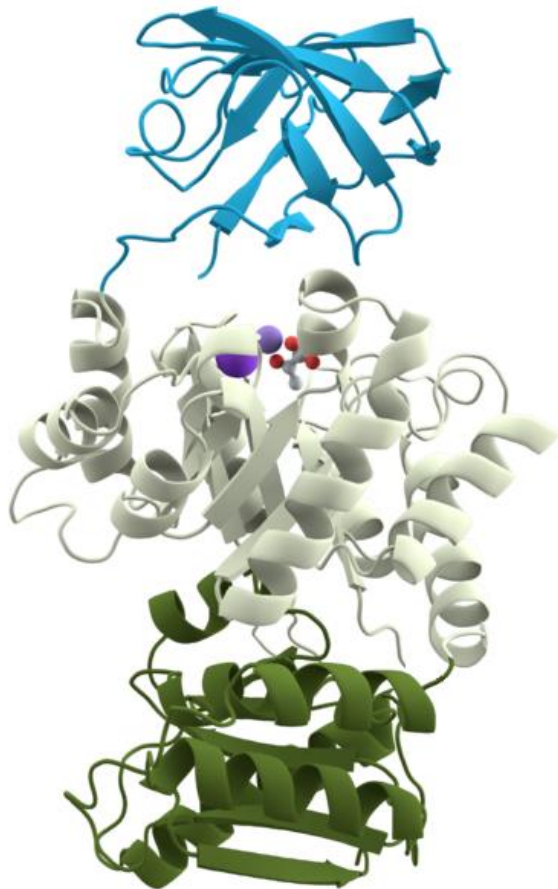


PFD**F**SPV**C**ATEL

From: Peroxiredoxins are conserved markers of circadian rhythms. Nature 485, 459–464 (24 May 2012)

Motivations for sequence alignment

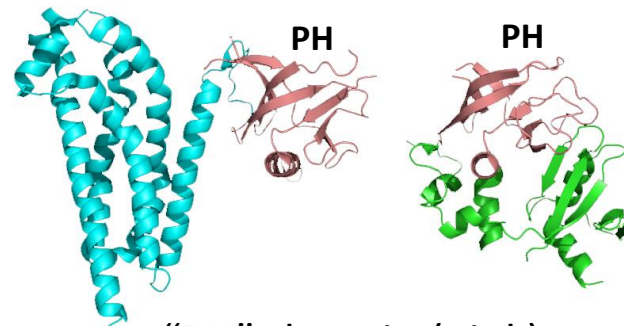
3) to identify and characterize “protein domains”



Pyruvate Kinase

definition:

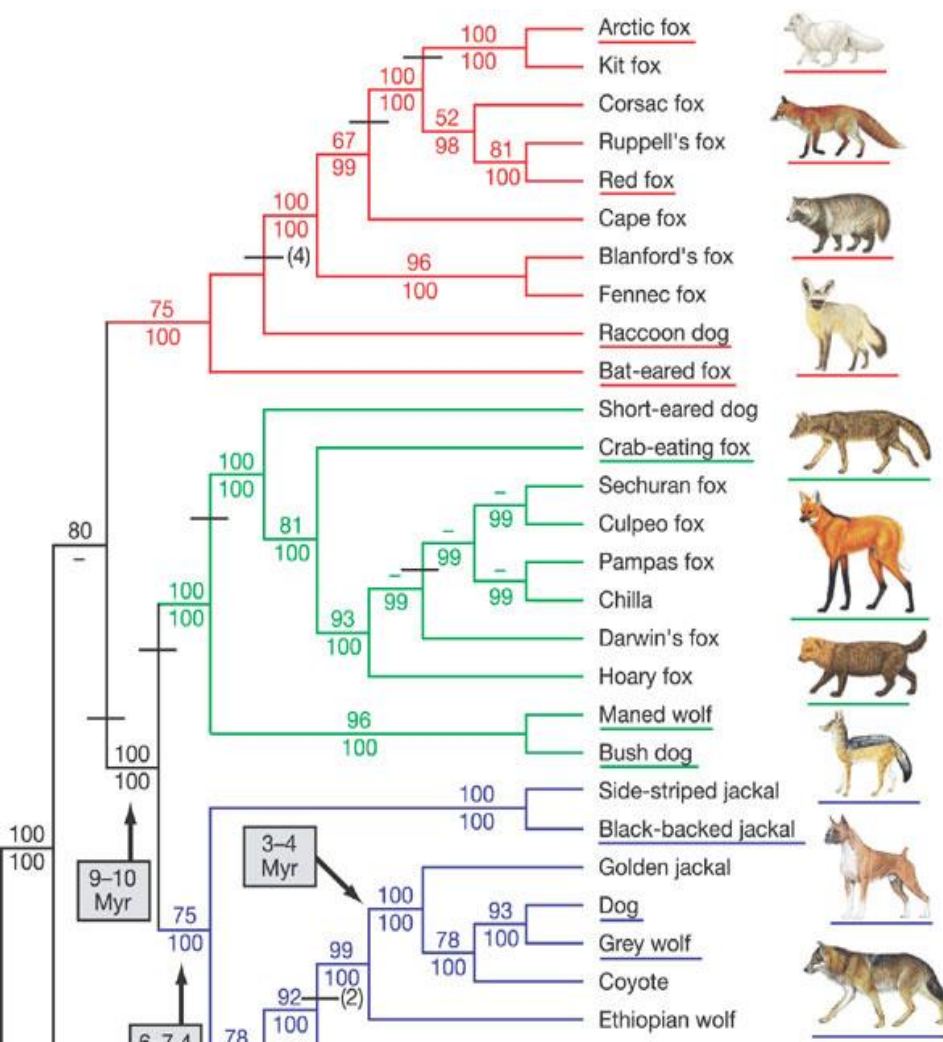
“parts of proteins that can evolve, function, and exist independently of the rest”



“PH”-domain (pink);
occurring in two different proteins

Motivations for sequence alignment

4) to make phylogenetic inferences (“trees”)



canine 1
canine 2
canine 3
[...]

IRMSLRGK	-	AVVLMGKNTMMRKAIRGHLENN	--	PALE
IRMSLRGK	-	AVVLMGKNTMMRKAIRGHLENN	--	PALE
IRMSLRGK	-	AVVLMGKNTMMRKAIRGHLENN	--	PALE
IRMSLRGK	-	AVVLMGKNTMMRKAIRGHLENN	--	PALE
IRMSLRGK	-	AVVLMGKNTMMRKAIRGHLENN	--	PALE
IRMSLRGK	-	AVVLMGKNTMMRKAIRGHLENN	--	SALE
IRLSLRGK	-	AVVLMGKNTMMRKAIRGHLENN	--	PALE
IRLSLRGK	-	AIVLMGKNTMMRKAIRGHLENN	--	PALE
IRTSLRGL	-	AVVLMGKNTMMRKAIRGHLENN	--	PQLE
IRKSIRGI	-	GAVLMGKKTMIKRVIRDLADSK	--	PELD
IRKSIRGI	-	GAVLMGKKTMIKRVIRDLADSK	--	PELD
VRKSLRGK	-	ATILMGKNTIRIATLKKNLQAV	--	PQIE
IRKKLRGK	-	ADIKVTKNLNFNIALKNAG	----	YDTK
IRKKMRGM	-	AEIKVTKNTLFGIAAKNAG	----	LDVS
IRKKLRGK	-	ATIKVTKNTLFKIAAKNAG	----	IDIE
VRKKLWKK	-	YPMMVAKKRIILRAMKAAGLE	----	LDDN
YRYRLRRY	-	GVIKIIPKPTLFKIAFTKVYGG	----	IPAE
IRRDLDKV	-	AVLVSRNTLTERRALNQLG	----	ETIP
IRRDLDKV	-	AVLVSRNTLTERRALNQLG	----	ESIP
IRREFRGK	-	AEIKVYKNTLLERLALDALG	----	GDYL
IRAKLRERD	-	TIIRMSRNTLMRIALEEKLDER	--	PELE
MRQTLRDS	-	ALIRMSKKTLLISLAEKAGREL	--	ENVD
IRDKIR	-	GTMTLKMSRNTLIERAIKEVAEETGNPEFA	--	ENVD
IRDKIR	-	DQMTLKMSRNTLIKRAVEEVAEETGNPEFA	--	ENVD
IRDKIR	-	DKVKLRMSRNTLIIRALKEAAEELNPKLA	--	ENVD
MRRLIRENG	-	GLLRVSRNTLIELAIKKAQELGKPELE	--	ENVD
MRRLIRENG	-	GLLRVSRNTLIELAIKKAQELGKPELE	--	ENVD
MRRLIRENG	-	GLLRVSRNTLIELAIKKAQELGKPELE	--	ENVD
MRDKLR	-	GKALLRVSRNTLIELAIKRAQELGQPELE	--	ENVD
MRDLHGT	-	AELRVSRNTLLELALDDVD	----	DGLE
MRRELHGS	-	AAVRMSRNTLVNRLDEVN	----	DGFE
MRRLHGO	-	AALRMSRNTLLVRALEEAG	----	DGLD
IRGKNRGK	-	INLVIKKTLTLFKALENLGD	----	EKLS

How it's done ...

Chymotrypsin VKKTMVCAGG-DGVISACNGDSGGPLNCQLENGSWEVFGIVSFGSRRGC [...]
 + M C G +G +C GDSGGP+ C NG + G+VS+G GC
 Trypsinogen ITSNMFCVGFLEGGKDSCQGDSSGGPVVC---NGQLQ--GVVSWGD--GC [...]

NCQLENGSWEV
 C NG +
 VC---NGQLQ--

=> “substitution matrix” (learned from structures)

- Why not align “QL” instead of “NG” ??
- What does the “+” mean ?
- How “good” is my alignment?

	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

Pairwise Alignment

a) BLAST: quick and dirty

VKKTMCAGG-DGVISACNGDSGGPLNCQLENGSWEVFGIVSF
GGP

1) identify word-matches
(use indexing)

← VKKTMCAGG-DGVISACNGDSGGPLNCQLENGSWEVFGIVSF →
+ M C G +G +C GDSGGP+ C NG + G+VS+
ITSNMFCVGFLEGGKDSCQGDSGGPVVC---NGQLQ--GVVSW

2) extend into an “HSP”
(high-scoring sequence pair)

b) Dynamic Programming – correct and slow

	A	T	T	G	G	A	G	T	C	G
A	1	⇒0	⇒-1	⇒-2	⇒-3	⇒-4	⇒-5	⇒-6	⇒-7	⇒-8
T	↓0	↘2	↖1	→0	→-1	→-2	→-3	→-4	→-5	→-6
C	↓-1	↓1	↘2	→1	→0	→-1	→-2	→-3	↖-3	↖-4
T	↓-2	↘0	↖2	→2	→1	→0	→-1	↖-1	↖-2	↖-3
C	↓-3	↓-1	↘1	→2	→2	→1	→0	→-1	↖0	↖-1
G	↓-4	↓-2	↘0	→2	→3	→2	→2	→1	↖0	↖1

ATTGGAGTCG

83%

ATC-----TCG

OR

ATTGGAGTCG

83%

AT-----CTCG

explore all possible paths
(but excluding obvious
dead ends)

Multiple Alignment

```

      10      20      30      40      50      60      70      80      90     100     110     120     130
1 -----MALEKSLVRLLLLVLIL-----LVLGWVQPSLGKE-SRAKKFQRQHMDSDSSPSSSS-----TCNQMMRR--RNMTQ 65
1 -----MVP-KLFTSQICLLLLLGL-----LAVEGSLHVKPPQFTWAQWFETQHINMTSQ-----QCTNAMQV--INNYY 61
1 -----MVP-KLFTSQICLLLLLGL-----MGVEGSLHARPPQFTRAQWFAIQHISLNPP-----RCTIAMRA--INNYY 61
1 -----MALEQRTHSLLLLLLLTL-----LGLGLVQPSYQQD-GMYQRFLRQHVFPEETGGSD-----RYCNLMQMQR--RKMTL 64
1 -----MVMG-----LGVLLLLVFV-----LGLGLTPPTLAQDNSRYTHFLTCHYDAKPPQGRDD-----RYCESIMRR--RGLTS 61
1 -----MVLCFP-LLLLLLVL-----WGPVCPHAWPKRLTKAHWFEIQHIQPSPL-----QCNRAMSG--INNYY 57
1 -----MAPARAGFCPLLLLLLLGL-----WVAEIPVSAKPKGMTSSQWFKIQHMQPSQ-----ACNSAMKN--INKHT 62
1 -----MAPARAGCCP-LLLLLLGL-----WVAEVLVRAKPKDMTSSQWFKTCHVQPSQ-----ACNSAMSI--INKYY 61
1 MMRTLITTHPLPLLPLPQQQLQLVQFQEVDTDFDFPEEDKKEEFEECLEKFFSTGPARPPTKEKVKRRVLEPG-----MPLNHIEYCNHEIMG-KNVYY 95
1 -----MKLNLVQIFFMLMLLLGLGMGLGLSLHMAVLEESDQPLNEFWSSDSQDKAEATEEGDGTQTTETLVLSNKEVVQPGWPEDPILGEDEVGGNKMLRASALFQSNKDYLRLDQTDRECNDDMAHK-MKEPS 131
1 -----METFPLLLLSLGLVLAEESESTMKIIEKEFTDEEMQYDMAKSGQEKQTIEILMNPILLVKN-----TSLSMKDDMSSTLLTFRSLHYNDPKGNSSGNDKECCNDMTVWRKVSEAN 111
1 -----MIIMVLIIFLVLLF-----WENEVND EAVMSTLEHLHVDYPQNDYFVPAR-----YCNHMIIQRVIREPD 59
1 -----MAPAVTRLFLQLVLG--PTLVMDIKMQIGSRNFYTLSDYPRVNYPKGFRG-----YCNGLMSYMRGKMQN 65
1 -----MALK-SLVLLSLLVLVL-----LLV-RVQPSLGKE-TAAAKFERQHMDSSTSAASSS-----NYCNQMMS--RNLT 63

```

Combinatorial Explosion: very many possible solutions

Complexity: $O(\text{alignment_length}^{\text{number_seqs}})$

=> an NP-complete problem !!

not feasible

Multiple Alignment

TCOFFEE

PROBCONS

MUSCLE



SATé



MAFFT



PRANK

DIALIGN-TX

Quality of MSA: Benchmarking

has high quality alignments in this database



“BALiBASE”:

Benchmark Alignment Database

Hand-made multiple sequence alignments
Based on selected structural alignments

Structural Alignments
offer the best
benchmarks !

Welcome to BALiBASE 4

[download the whole benchmark by html](#)

Reference 1: variability, length

Reference 1: variability, length

Reference 2: orphans

Reference 3: sub-families

Reference 4: extensions

Reference 5: insertions

References 6,7,8: Repeat, Transmembrane, Circ. permutation

Reference 9: linear motifs

Reference 10: mixed

problem with an alignment : contact Julie Thompson

problem with the web site : contact Raymond Ripp

Practical Session: Source of your Sequences



**Dental Calculus of
Medieval Mummy**

(with kind permission of
Christina Warinner, UZH)

