Regaining our capacity for surprise Jeff Elhai, Arnaud Taton, JP Massar, Jeff Shrager

The study of microbes has been tied to observation, through microscopy and the careful scrutiny of macroscopic characteristics. The study of the molecular biology of microbes is similarly connected to observation, for example electron micrographs of adenoviral RNA-DNA hybrids leading to the discovery of introns (1,2). The most basic discoveries in biology, regardless of discipline, are generally associated with the word "surprise" and a deep connection between the observer and the object of study, permitting a radical new interpretation of the surprising observation.

Why state the obvious?

Because the close connection between the biologist and the phenomenon under study is breaking down. We are paying for the rapid progress made possible by the new wealth of biological information by a reduced connection with unprocessed data and the consequent reduced possibility for productive surprises. We gain in the short term. In the long run, we're in for trouble.

The problem is that the information we have been gifted with – genomes, microarrays, data from other high-throughput experiments – demand computation in order to make sense of it, and few biologists are comfortable with computation. So what? Few biologists are comfortable with the physics of optics, yet we use microscopes. Few of us can build a spectrophotometer, yet we can still understand what a spectrum may tell us. The relationship of biologists with bioinformation is fundamentally different, however. Biologists using these older tools of biology could see the object of their study to the extent that it was technically possible to be seen. Perhaps a spectrophotometric scan is automated and provides intensities at preset wavelengths. If we're suspicious, we are fully capable of looking at the full spectrum or even painfully examining hundreds of specific wavelengths if we choose to. Few biologists are equally capable of examining raw genomic information or any other mass data, because that would require going outside of fixed computational tools and creating a new set of directions. That would require control over the process, i.e. programming the computer.

As a result, most of the growing number of biologists that use mass information confine themselves to the few computational tools, like Blast and Clustal, that have well developed user interfaces that cater to the noncomputational biologist. We gain comfort from this ease of use and the familiarity that comes from repetition, but that layer of comfort is thin. As soon as the tool begins to behave in ways we consider out of bounds, we are thrown back into confusion, with no means by which to steady our gaze and find new order.

The need for computational agility: A scenario

Suppose, for example, that you are interested in a protein, Asr1156 (GenBank NP_485199), from the cyanobacterium *Anabaena* PCC7120, that is induced during differentiation and annotated as "hypothetical". Blast returns the result that it is similar to the C-terminus of proteins from all or nearly all cyanobacteria with completed genomes. The protein is usually annotated as "an anti-sigma factor antagonist", a type of protein that might indeed be interesting with respect to differentiation. On most days, that would be the end of it, but suppose that today you expend the effort necessary to obtain the sequences of the similar protein and to align them, getting the

Δ								
	·	_20	_30	40	50	60	70	80
5-WH7805_02012	MH <mark>PGG</mark> SV	PITELQRL'	TV <mark>S</mark> LR <mark>gg</mark> f <mark>e</mark> q	K <mark>gg</mark> clvfhf	TGQLDAFSEK	QFLS <mark>YVAD</mark> VL	KANAS <mark>s</mark> tvl <mark>d</mark>	L <mark>s k i d</mark> f l <mark>d s s g</mark> l
p-SynWH7803_1062	2		· · <u>·</u> · · <u>· · · · ·</u> ·	• • • • • MF H F	TGQLDAYSEK	QFLS <mark>YVTD</mark> VL	KANAS <mark>PT</mark> VL <mark>D</mark>	L <mark>SKID</mark> FL <mark>DSSG</mark> L
5-RS9917_11490		· · · · · · L	TV <mark>S</mark> LR <mark>gg</mark> f <mark>e</mark> q	RDRCLVFHF	TGQLDAYSEK	QFLA <mark>YTGD</mark> VL	Q T N K S A V V V D	L <mark>skidfidssg</mark> l
5-RS9916_28814				· · · · · LF HF	TGQLDAYSER	QFLTYVSDVL	KANTSSLVID	L <mark>SKIDFMDSSG</mark> L
5-PMT0652				MFQF	TGQLDAYSEK	QFITYVEDVL	MANAAPLVID	LNKIDFLDSSGL
5-P9303_15781		MQRL	TVSLRGGFER	ROGCLVFOF	TGULDAYSEK	QFITYVEDVL		LNKIDFLDSSGL
5-P700/46		PITELUKL		KKGCLVFYF	TOOLDAYSEK		SSNKLPIVLD	LTKIDFIDSSGL
5-09211205337 = NAT24 09294				MFYF	TGOLDAYSEK		ASNUL PMVID	LENIDEIDECAL
5-RMM2A 0260					TGOLDAYSEK		ASNOLSVVID	
5				REDCLLESE	TGOLDAYSDK		ASNGESVVID ASNKRAVIID	L THIDE IDSSGL
n-SvaRCC307_1330	RICSSK <mark>g</mark> SR	TITDIORI	TVSIRAGEEP	ODOCLLEHE	TGOLDAYSER	DERDEIGDNA	PSGI KPLVID	LSKIDEIDSSGL
5-Cw?2516			- VSLRGTHEV	KGNYDIFRL	TGLLDAFSEP	VESKVVKOYI	EEGPYNIILV	
5-00110203529			- MSLRGTHEV	KGNCQIFRL	TGLLDAFSEP	VERKVVGQYI	DEGPYDIVLV	LAQIDFIDSSGL
p-Asr1156							MIL <mark>D</mark>	L <mark>SQIDFVDSSG</mark> L
5-Ava4411	SVHFNDEE <mark>g</mark>	ITAE <mark>P</mark> ENE	TV <mark>S</mark> LR <mark>G</mark> TR <mark>E</mark> V	RDNYQLFRL	TGLLDAFSEP	TFRKVL <mark>gg</mark> ki	DE <mark>gp</mark> khiil <mark>d</mark>	L <mark>SQIDFVDSSG</mark> L
5-79414?03196			- M <mark>s</mark> lr <mark>g</mark> tr <mark>e</mark> v	R D N C Q L F R L	. TGLLDAFSEP	TFRKVL <mark>g</mark> nki	DE <mark>gp</mark> khiil <mark>d</mark>	L <mark>S Q I D</mark> F V <mark>D S S G</mark> L
5-NpR1192					L L <mark>D</mark> A F <mark>S E P</mark>	TFRKVL <mark>g</mark> ski	EE <mark>GP</mark> KHIIL <mark>D</mark>	L <mark>S Q I D</mark> F I <mark>D S S G</mark> L
p - ∤8106?28646		M	T V <mark>S</mark> L R <mark>G</mark> N R <mark>D</mark> V	KDKYQ <mark>lf</mark> rl	. TGLLDAFSEA	TFRKVV <u>sk</u> ci	E <mark>ggp</mark> tnivl <mark>d</mark>	L <mark>S Q I D</mark> F V <mark>D S S G</mark> L
p-Synpec7942_2121	1 MTTLEDT	AIRETLAL	TV <mark>S</mark> LR <mark>G</mark> TV <mark>E</mark> V	R D D Y Q L F R L	. TGQLDAFSEP	T F R K V L <mark>g</mark> k <mark>s</mark> l	ED <mark>gp</mark> khail <mark>d</mark>	L <mark>skid</mark> fi <mark>dssg</mark> l
5-syc 1971	MTTLEDT	AIRETLAL	TV <mark>S</mark> LR <mark>G</mark> TV <mark>E</mark> V	RDDYQLFRL	. TGQLDAFSEP	TFRKVL <mark>G</mark> K <mark>S</mark> L	ED <mark>gp</mark> khail <mark>d</mark>	L <mark>skid</mark> fi <mark>dssg</mark> l
5-Sar1600		M	T V <mark>S</mark> L R <mark>G</mark> T L <mark>E</mark> V	KEKLQ <mark>IF</mark> R <mark>L</mark>	. TGLLDAFSEP	TFRKVVS <mark>DH</mark> I	EA <mark>gpp</mark> hvvl <mark>d</mark>	L <mark>S T I D</mark> F I <mark>D S S G</mark> L
5-CYA_0902				· · · · · <u>· · ·</u> · <u>·</u>		····MVTKYL	EE <mark>G</mark> S <mark>P</mark> NLILD	L <mark>T G I D</mark> F I <mark>D S S G</mark> V
5-CYB_2190	MVERRR	AI <mark>P</mark> E <mark>P</mark> LSL	TV <mark>S</mark> LR <mark>G</mark> SR <mark>E</mark> A	LENCQVFHL	TGLLDAFSEP	V F V K V V T K Y L	EE <mark>G</mark> SPNLILD	L <mark>T G I D</mark> F I <mark>D S S G</mark> V
o-gar4179						MMNKYI	EE <mark>GSP</mark> NIVLD	L SQ I D F V D S SG L
5-tsr0427						<mark>MI</mark> SK <mark>CF</mark>	DD <mark>GP</mark> ANVILD	LSKIEFIDSSGL
p-Tery_2311							MD	L <mark>s</mark> ki <mark>d</mark> fv <mark>dssg</mark> l
П								
В	90	100	110	120	130	140	150	160
В	90	100	110	120	130	140	150	160
B 5-SynW1027	90 - LRGETGPI	100 SEL Q.R.L.	110 T <mark>VSLRGGFEQ</mark>		130	140 OFMEYVVDVL	150 KASKL <mark>PAVID</mark>	160 L <mark>SKIDFLDSSGL</mark>
B SynW1027 WH7805_02012	90 - LRGETGPI MHPGGSVPI	100 SEL QRL TEL QRL	110 TVSLRGGFEQ TVSLRGGFEQ	120 DNGCLVFHF KGGCLVFHF	130 Tgoldaysek Tgoldaysek	140 DEFMEYVVDVL OFLSYVADVL	150 Kaskl <mark>pavid</mark> Kanass tvl d	160 LSKIDFLDSSGL LSKIDFLDSSGL
B p-SynW1027 p-WH7805_02012 p-RS9916_28814 - RS9916_28814	90 • LRGETGPI MHPGGSVPI NKPWGPGPI	100 SEL QRL TEL QRL TEL QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ	120 NGCLVFHF KGGCLVFHF RDNCLLFHF	130 TGQLDAYSEK TGQLDAFSEK TGQLDAYSEK	140 QFMEYVVDVL QFLSYVADVL QFLSYVADVL	150 Kaskl <mark>pavid</mark> Kanass t vld Kantsslvid	160 LSKIDFLDSSGL LSKIDFLDSSGL LSKIDFMDSSGL
B p-SynW1027 p-WH7805_02012 p-RS9916_28814 p-RS9917_111490 - P=07406	90 - LRGETGPI MHPGGSVPI NKPWGPGPI GSG <mark>S</mark> I PSIDGCGPI	100 SEL - QRL TEL - QRL TEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ	120 NGCLVFHF KGGCLVFHF RDNCLLFHF RDRCLVFHF	130 TGQLDAYSEK TGQLDAFSEK TGQLDAYSEK TGQLDAYSEK	140 CFMEYVVDVL CFLSYVADVL CFLSYVADVL CFLSYVSDVL CFLAYTGDVL CFLSYTGDVL	150 Kaskl <mark>pavid</mark> Kanass t vld Kantsslvid QTNKSAVVVD SSNUL <mark>B</mark> UVID	160 LSKIDFLDSSGL LSKIDFLDSSGL LSKIDFMDSSGL LSKIDFIDSSGL
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Pro0746 Pro0746	90 • LRGETGPI MHPGGSVPI NKPWGPGPI • • • • GSG <mark>S</mark> I RSIRGFGPI PSTPARP	100 SEL - QRL TEL - QRL TEL - QRL TEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER	120 CNGCLVFHF KGGCLVFHF RDNCLLFHF RDRCLVFHF KKGCLVFYF	130 TGQLDAYSEK TGQLDAFSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK	140 CIFLSYVADVL CIFLSYVADVL CIFLSYVSDVL CIFLAYTGDVL CIFLAYTGNDVL CIFTDFVNDVL	150 KASKL <mark>PAVID</mark> KANASS T VID KANTSSLVID QTNKSAVVVD SSNKLPIVLD SSNKLPIVLD	160 LSKIDFLDSSGL LSKIDFLDSSGL LSKIDFMDSSGL LSKIDFIDSSGL LTKIDFIDSSGL
B 	90 - LRGETGPI MHPGGSVPI NKPWGPGPI GSGSI RSIRGFGPI RSIRGFGPI RSTRRAKPI	100 SEL - QRL TEL - QRL TEL - QRL TEL - QRL TEL - QRL KDL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER	120 GCLVFHF KGGCLVFHF RDRCLVFHF KGCLVFYF KGCLVFYF RKGCLVFYF	130 TGQLDAYSEK TGQLDAFSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK	140 COFMEYVVDVL COFLSYVADVL COFLSYVSVL COFLAYTODVL COFTDFVNDVL COFTDFVNDVL	150 KASKL <mark>PAVID</mark> KANASS T VLD KANTSSLVID QTNKSAVVVD SSNKL <mark>PIVLD</mark> SANQL P MVID ASNOLSWVID	160 LSKIDFLDSSGL LSKIDFLDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LTKIDFIDSSGL LTKIDFIDSSGL
B 	90 - L R G E T G P I MH P G G S V P I N K P WG P G P I R S I R G F G P I R S T R R A K P I G T N P I	100 SEL - QRL TEL - QRL TEL - QRL TEL - QRL TEL - QRL KDL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER	120 0 NGCLVFHF KGGCLVFHF RDRCLVFHF KGCLVFHF KGCLVFYF 0 NGCLVINF	130 TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE	140 CGF LE YV DVL CGF LE YV ADVL CGF LA YT ODVL CGF TDF VN DVL CGF TDF VN DVF CGF TTY IN DVL FTTY IN DVL	150 KASKL <mark>PAVID</mark> KANASS T VLD KANTSSLVID QTNKSAVVVD SSNKL <mark>P</mark> IVLD SANQLPMVID ASNQLSVVID	160 LSKIDFLDSSGL LSKIDFLDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LTKIDFIDSSGL LTKIDFIDSSGL LSNIDFIDSCGL
B SynW1027 WH7805_02012 R59916_28814 R59917_11490 -Pro.0746 p9211?05337 NATL1_09291 PM102A_0260 PM10652	90 MHPGGSVPI NKPWGPGPI GSGSI RSIRGFGPI RSIRGFGPI GTNPI GTNPI IGTNPI	100 SEL - QRL TEL - QRL TEL - QRL TEL - QRL KDL - QRL TEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER	120 NGCLVFHF KGGCLVFHF RDRCLVFHF KKGCLVFHF KKGCLVFYF ONGCLVINF ONGCLVINF	130 TOQLDAYSEK TOQLDAYSEK TOQLDAYSEK TOQLDAYSEK TOQLDAYSEK TOQLDAYSEK TOQLDAYSEK TOQLDAYSEK	140 CG F ME YVV DVL CG F LS YVA DVL CG F LT YVS DVL CG F TD F VN DVL CG F TT Y I N DVL CG F TT Y I N DVL CG F TT Y I N EVL CG F TT Y I N EVL CG F TT Y I N EVL CG F TT Y I N EVL	150 KASKL PAVID KANASSTVLD KANTSSLVID QTNKSAVVVD SSNKL PIVLD SSNKL PIVLD SANQLSVVID ASNQLSVVID ASNQLSVVID	160 LSKIDFLDSSG LSKIDFLDSSG LSKIDFIDSSG LSKIDFIDSSG LTKIDFIDSSG LTKIDFIDSSG LSNIDFIDSCG LSNIDFIDSCG LNKIDFIDSSG
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Pro0746 Pro0746 Pa0745337 NATL1_09291 PMM2A_0260 PMM0652 PM0652 PM0652 P303 15781	90 - LRGETGPI MHPGGSVPI NKPWGPGPI GSGSI RSIRGFGPI RSTRAKPI GTNPI GTNPI IASRGTGPI	100 SEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGSNQ TVSLRGGSER TVSLRGGFER		130 TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH	140 CGF ME YVVD VL CGF LSYVAD VL CGF LSYVAD VL CGF LTYVSD VL CGF TDF VND VF CGF TTY IND VL CGF TTY IND VL CGF TTY VED VL CGF TTYVED VL	150 KASKLPAVID KANASSTVID KANTSSLVID QTNKSAVVVD SSNKLPIVLD SSNQLSVVID ASNQLSVVID ASNQLSVVID MANAAPLVID LANAAPLVID	160 LSKIDFLDSSGL LSKIDFMDSSGL LSKIDFMDSSGL LSKIDFIDSSGL LTKIDFIDSSGL LTKIDFIDSGGL LSNIDFIDSGGL LSNIDFIDSGGL LNKIDFLDSSGL
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Pro0746 p9211?05337 NATL_1_09291 PMN2A_0260 PMT0652 PMT0652 PMT0652 PMT0652 PMT0652 PMT0652 SwnRCC307_1330	90 - LRGETGPI MHPGGSVPI NKPWGPGPI GSGS RSTRAKPI GTNPI IASRGTGPI IASRGTGPI CSSKGSRTI	100 SEL - ' QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER		130 TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK	140 G F ME Y V V D V L G F L S Y V A D V L G F L A Y T G D V L G F T D F V N D V L G F T T Y I N D V L G F T T Y I N E V L G F T T Y V E D V L G F T T Y V E D V L G F T T Y C D V L G F T T Y C D V L G F T T Y C D V L G F T D T O D N A	150 KASKLPAVID KANASSTVID KANTSSLVID QTNKSAVVVD SSNKLPIVLD SANQLSVVID ASNQLSVVID ASNQLSVVID MANAAPLVID LANAAPLVID PSGLKPLVID	160 LSKIDFLDSSGL LSKIDFMDSSGL LSKIDFMDSSGL LSKIDFIDSSGL LTKIDFIDSGGL LSNIDFIDSGGL LSNIDFIDSGGL LNKIDFLDSSGL LNKIDFLDSSGL LNKIDFLDSSGL
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Pm0746 p92117205337 NATL_09291 PM170652 PM07652 PM07652 PM07651 SynRCC307_1330 WH5701_08484	90 . L R C E T C P I MHPGGSVPI NKPWGPGPI GSG SI RSIRGFGPI RSIRGFGPI I ASRGTGPI I ASRGTGPI I ASRGTGPI I ASRGTGPI C S L A R G T I	100 SEL - QRL TEL - QRL TEL - QRL TEL - QRL KDL - QRL TEL - QRL TEL - QRL TEL - QRL TEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER	120 NGCLVFHF RDNCLVFHF RDNCLVFHF KGCLVFYF QNGCLVFYF QNGCLVFYF QNGCLVFQF RQGCLVFQF RQGCLVFQF RQGCLLFSF	130 TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK	140 G F L S Y V D V L G F L S Y V A D V L G F L T Y V D V L G F T D F V N D V L G F T T F V N D V L G F T T Y I N E V L G F T T Y V E D V L G F T T Y V E D V L G F T D F I D T E Q R G F R D F I E Q R	150 KASKLPAVID KANASSTVID KANTSSLVID GINKSAVVVD SSNKLPIVLD SSNQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID GSLKRAVLID GSDKRAVLID	160 LSKIDFLDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LTKIDFIDSSGL LSNIDFIDSGGL LSNIDFIDSGGL LNKIDFLDSSGL LNKIDFIDSSGL LSKIDFIDSSGL
B 	90 - L R G E T G P I MH PG G S V P I NK PWG PG S V P I R S I R G F G P I R S I R G F G P I R S T R A K P I C S L A R G T G P I I A S R G T G P I I A S R G T G P I I A S R G T G P I C F L A R A G T I S T N V K E E A I	100 SEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER	120 CLVFHF KGGCLVFHF RDRCLVFHF KGCLVFYF RGCLVFYF NGCLVINF RGCLVINF RGCLVFQF RGCLVFQF RGCLVFQF RGCLVFQF RGCLLFSF KGNYCLFSL	130 TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK	140 COFME YVVDVL COFME YVADVL COFTDFVNDVL COFTDFVNDVL COFTTVNDVF COFTTYNDVF COFTTYNDVL COFTTYN	150 KASKL PAVID KANASSTVD KANTSSVVD QTNKSAVVVD SSNKL PIVLD SANQL PWVD ASNQLSVVID ASNQLSVVID ASNQLSVVID LANAAPLVID CANAAPLVID SSDKRAVLLD EEGPYNILL	160 LSKIDFLDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LTKIDFIDSSGL LTKIDFIDSSGL LSNIDFIDSCGL LNKIDFLDSSGL LNKIDFLDSSGL LNKIDFIDSSGL LSKIDFIDSSGL LSKIDFIDSSGL
B SynW1027 WH7805_02012 R59916_28814 R59917_11490 Pro0746 99211705337 NATL1_09291 PMN2A_0260 PM70652 P9303_15781 SynRCC307_1330 WH5701_08484 Cw72516 OU10703529	90 H R G E T G P I M H P G G S V P I N K P WG P G P I R S I R G F G P I R S T R R A K P I G T N P I I A S R G T G P I I A S R G T G P I I A S R G T G P I C S S K G S R T I S T N V K E A I	100 SEL - QRL TEL - QRL TDL - QRL PEL - HRL PEQ - LNL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGFER TVSLRGFER TVSLRGFER		130 TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE TGQLDAYSE	140 CG F LA YT O VL CG F LA YT O VL CG F LA YT O VL CG F T D F VN D VL CG F T T Y I ND VL CG F T T Y I ND VL CG F I T Y VE D VL CG F I	150 KASKL PAVID KANASS TVD KANTSSVVD SSNKL PIVLD SSNKL PIVLD SSNKL PIVLD SANQL SVVID ASNQLSVVID ASNQLSVVID MANAAPLVID LANAAPLVID PSGLKPLVID EEGPYNILV DEGPYNIVLV	160 LSKIDFLDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LTKIDFIDSSGL LSNIDFIDSCGL LNKIDFIDSCGL LNKIDFLDSSGL LNKIDFLDSSGL LNKIDFLDSSGL LSKIDFIDSSGL
B SynW1027 WH7805_02012 R59916_28814 R59917_11490 Pro0746 Pro0746 Pro0746 Pro0746 Pro0746 Pro0745 Pro0745 Pro0745 Pro0745 Pro0745 Pro0745 SynRCG307_1330 WH5701_08484 CW72516 CW72516 Dro07529 Asr1756	90 - LRGETGPI MHPGGSVPI NKPWGPGPI GSGSI RSTRAKPI GTNPI GTNPI IASRGTGPI IASRGTGPI CFLARAGTI STNVKEEAI STNVKEEAI	100 SEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGTHEV		130 TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH TGQLDAYSEH	140 CGF ME YVVD VL CGF L TYVSD VL CGF L TYVSD VL CGF L TYVSD VL CGF T TY I ND VL CGF T TY I ND VL CGF T TYVED VL CGF T TYVED VL CGF T TYVED VL GF T TYVED VL TYVED VL GF T TYVED VL GF T TYVED VL GF T TYVED VL TYVED VL GF T TYVED VL GF T TYVED VL GF T TYVED VL TYVED VL GF T TYVED VL GF	150 KASKLPAVID KANASSTVID KANTSSLVID QTNKSAVVVD SSNKLPIVID SSNQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID QSDKRAVID CANAAPLVID SSUKPLVID GSDKRAVILD EEGPYNILV DEGPYDIVLV DEGPYDIVLV	160 LSKIDFLDSSGL LSKIDFMDSSGL LSKIDFMDSSGL LSKIDFIDSSGL LTKIDFIDSSGL LSNIDFIDSGGL LSNIDFIDSGGL LSNIDFIDSGGL LNKIDFLDSSGL LNKIDFLDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LSGIDFIDSSGL
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Pro0746 p921105337 NATL1_09291 PMT0652 PMT0652 P9303_15781 SymRC6307_1330 WH5701_08484 Cw72516 Cw72517 	90 - LRGETGPI MHPGGSVPI NKPWGPGPI GSGS RSTRAKPI GTNPI IASRGTGPI IASRGTGPI CSSKGSRTI CFLARAGTI STNVKEEAI STNVKEEAI HFNDEEGII	100 SEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGTHEV TVSLRGTHEV TVSLRGTREV		130 TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGLDAFSEF TGLLDAFSEF	140 G F ME Y V V D V L G F L S Y V A D V L G F L S Y V A D V L G F T D F V N D V L G F T T Y I N D V L G F T T Y I N E V L G F T T Y V E D V L G F T T Y E D V L G F T T T Y E D V L G F T T T T T T T T T T T T T T T T T T	150 KASKLPAVID KANASSTVID KANTSSLVID QTNKSAVVVD SSNKLPIVLD SSNQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID BSGLKRAVLLD EEGPYNIILV DEGPKHIILD DEGPKHIILD	160 LSKIDFLDSSGL LSKIDFMDSSGL LSKIDFMDSSGL LSKIDFIDSSGL LSKIDFIDSGGL LTKIDFIDSGGL LSNIDFIDSGGL LSNIDFIDSGGL LSNIDFIDSGGL LNKIDFLDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LSGIDFVDSSGL
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Ro0746 p9211?05337 PMV2A_0260 PMT0652 P9303_15781 SynRCC307_1330 WH5701_08484 CW?2516 0110?03529 Ava4411 Ava4411 Ava4411	90 - LROETOPI MHPGGSVPI NKPWGPGPI GSGSI RSIRGFGPI GTNPI IASRGTGPI IASRGTGPI IASRGTGPI IASRGTGPI STNVKEEAI STNVKEEAI STNVKEEAI HFNDEEGII	100 SEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGTHEV TVSLRGTHEV TVSLRGTREV	120 NGCLVFHF RDNCLLFHF RDNCLVFHF KGCLVFYF QNGCLVFYF QNGCLVFYF QOCLVFQF RQGCLVFQF RQGCLVFQF RQGCLVFQF RQGCLLFSF KGNCQIFRL RDNYQLFRL RDNYQLFRL RDNCQLFRL	130 TGQLDAYSEK	140 G F L Y V V D V L G F L Y V S D V L G F L Y V S D V L G F T D F V N D V L G F T T Y I N E V L G F T T Y I N E V L G F T T Y V E D V L G F T T Y V E D V L G F T T Y V E D V L G F T T F I T E Q R V F S K V V G Q Y I T F R K V L G G K I T F R K V L G K K I	150 KASKLPAVID KANASSTVID QTNKSAVVVD SSNKLPIVID ASNQLSVVID ASNQLSVVID ASNQLSVVID MANAAPLVID CANAAPLVID GSDKRAVID EEGPYNIILV DEGPKHIILD DEGPKHIILD	160 L S K I D F L D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L T K I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S K I D F I D S S G L L S L D F I D S S G L L T H I D F I D S S G L L A Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L
B 	90 - LRGETGPI MHPGGSVPI NKPWGPGPI GSGSI RSIRGFGPI GTNPI IASRGTGPI IASRGTGPI IASRGTGPI CSLARAGTI STNVKEEAI STNVKEEAI HFNDEEGII HFNDEEGII	100 SEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGFER TVSLRGTHEV TVSLRGTREV TVSLRGTREV TVSLRGTREV	120 C N G C L V F H F K G G C L V F H F R D N C L V F H F R C L V F Y F C N G C L F F R R D N C Q L F R L R D N C Q L F R L	130 TGQLDAYSEK	140 GF F ME YVVD VL GF F S VVD VL GF F T YVD VL GF T D F VND VL GF T T F VND VL GF T T YVED VL GF T T YVED VL GF T T YVED VL GF T T F T E QR VF S KVVG QYI T F R KVL G KI T F R KVL G KI	150 KASKL PAVID KANASSTVD KANTSSVVD SSNKL PIVD SSNKL PIVD SSNKL PIVD MANAAPLVID ASNQLSVVID ASNQLSVVID CANAAPLVID	160 LSKIDFLDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LSKIDFIDSSGL LSNIDFIDSSGL LSNIDFIDSSGL LNKIDFIDSSGL LSQIDFIDSSGL LSQIDFVDSSGL LSQIDFVDSSGL LSQIDFVDSSGL LSQIDFVDSSGL LSQIDFVDSSGL
B 	90 - L R G E T G P I MH PG G S V P I NK PWG PG G V I R S I R G F G P I R S I R G F G P I I A S R G T G	100 SEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGFEV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV	120 C N G C L V F H F K G C L V F H F R D N C L V F H F R C L V F Y F C N G C L V F F C N G C L V F Y F C N G C L V F F C N G C L F F C N G C L F C	130 TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAYSEK TGQLDAFSEF TGLLDAFSEF TGLLDAFSEF TGLLDAFSEF TGLLDAFSEF TGLLDAFSEF	140 C F ME YVVD VL C F L YVA D VL C F L YVS D VL C F T D F VN D VL C F T D F VN D VL C F T T Y I N D VL C F T T YVE D VL C F T YVE D VL C F T T T T T T T T T T T T T T T T T T	150 KASKL PAVID KANASSTVD KANTSSVVD SSNKL PIVLD SSNKL PIVLD SANQL SVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID MANAAPLVID CANAAPLVID BSGLKPLVID CEGPYNILL DEGPKHILD DEGPKHILD EGGPTNIVLD	160 L S K I D F L D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L T K I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L N K I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S I D F I D S S G L L S I D F I D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Pro0746 Pro0746 Pro0746 Pro0746 Pro0746 Pro0745 Pro0745 Pro0745 Pro0745 Pro0745 Pro0745 Pro0745 Pro0745 Pro0745 SynRCC307_1330 WH5701_08484 CW22516 Avr24111 N9414703196 N9414703196 N8414703196 N8414703196 N8414703196 N8414703196 N8414703196 N8414703196	90 - LRGETGPI MHPGGSVPI NKPWGPGPI GSGS RSIRGFGPI RSTRAKPI GTNPI GTNPI IASRGTGPI IASRGTGPI CFLARAGTI STNVKEAI STNVKEAI HFNDEEGII HFNDEEGII RFNHEEGII RFNHEEGII EGAI	100 SEL - QRL TEL - QRL NL A QR PEQ - LNL A EP - LNL A EP - LNL A EP - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGTEV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV		130 T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G L D A F S E H	140 G F ME Y V V D V L G F L A Y T O D V L G F L A Y T O D V L G F L A Y T O D V L G F T T Y I N D V L G F T T Y V E D V L G F T T T Y V E D V L G F T T T T T T T T T T T T T T T T T T	150 KASKL PAVID KANASSTVID KANTSSLVID QTNKSAVVVD SSNKL PIVLD SSNQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID DANAAPLVID CANAAPLVID SSDKRAVLLD COPYNILV DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD	160 L S K I D F L D S S G L L S K I D F MD S S G L L S K I D F MD S S G L L S K I D F I D S S G L L T K I D F I D S S G L L S N I D F I D S C G L L S N I D F I D S C G L L S N I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S Q I D F I D S S G L L S Q I D F I D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F I D S S G L L S Q I D F I D S S G L L S Q I D F V D S S G L L S Q I D F I D S S G L L S Q I D F I D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S K I D F V D S S G L
B SynW1027 WH7805_02012 R59916_28814 R59917_11490 Pro0746 p921105337 NATL1_09291 PMT0652 PMT06	90 - L R G E T G P I MH P G S V P I NK PWG P G P I G S G S I R S T R R A K P I G T N P I I AS R G T G P I C F L A R A G T I S T N V K E A I S T N V K E A I HF N D E G I I HF N D E G I I R F N H E E G I I T L T H E E G I I E G V I G C K A E E G V I	100 SEL - QRL TEL - QRL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGTEV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV	120 0 0 0 C L VF HF R D N C L LF HF R D N C L VF HF R K 0 C L VF YF 0 N 0 C L VF YF 0 N 0 C L VF NF 0 N 0 C L VF NF 0 0 0 C L VF NF R 0 0 C L FF R 0 0 C L FF	130 T G Q L D AY S E H T G L L D AF S E F T G L L D AF S E F	140 G F ME YVV D VL G F L YVS D VL G F L YVS D VL G F L YYS D VL G F T Y N D VL G F T T YV D VL G F T T YVE D VL G F T YVE D VL G F T T T T T T T T T T T T T T T T T T	150 KASKL PAVID KANASSTVID KANTSSLVID QTNKSAVVVD SSNKL PIVID SANQL PMVID ASNQLSVVID ASNQLSVVID ASNQLSVVID MANAAPLVID PSGLKPLVID EGGPYNIILV DEGPKHIILD DEGPKHIILD DEGPKHIILD DEGPKNILD DEGPKNILD DEGPKNILD DSGPKNILD	160 L S K I D F L D S S G L L S K I D F MD S S G L L S K I D F MD S S G L L S K I D F I D S S G L L T K I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S K I D F I D S S G L L S K I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S X I D F V D S S G L L S X I D F V D S S G L L S X I D F V D S S G L L S X I D F V D S S G L L S X I D F V D S S G L L S X I D F V D S S G L L S X I D F V D S S G L L S X I D F V D S S G L L S X I D F V D S S G L L S X I D F V D S S G L
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Ro0746 p9211?05337 PMV2A_0260 PMV0652 PM02A_0260 PMT0652 PM07052 SynRCC307_1330 WH701_08484 CW?2516 O110?03529 Ava4411 S41156 Ava4411 S41166?28646 Tery_2311 -Sar1600 Synpco7942_2121	90 - LROETOPI MHPGGSVPI NKPWGPGPI GSGSI RSIRGFGPI RSIRGFGPI IASRGTGPI IIII IIIII IIIIII IIIIII IIIIII IIIIII	100 SEL - QRL TEL - QRL NL PEQ - LNL AEP - LNL AEP - LNL AEP - LNL AEP - LNL AEP - LNL AEP - LNL	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGTHEV TVSLRGTHEV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV		130 T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A F S E F T G Q L D A F S E F T G Q L D A F S E F T G Q L D A F S E F T G Q L D A F S E F	140 GF L YVV DVL GF L YVV DVL GF L YVS DVL GF T DF VN DVL GF T DF VN DVL GF T T YV E DVL GF T T YVE DVL GF T T YVE DVL GF T F T VE DVL TF R KVL GK TF R KVL GK TF R KVVS KC TF R KVL G KS TF R KVL F R KVL F R K TF R KVL F R KVL F R K TF R KVL F R K F R K F R K F R K F R K F R K F R K F R K F R K F R K F R K F R K F R K F R K F R K F R K F R K F R K	150 KASKLPAVID KANASSTVID QTNKSAVVVD SSNKLPIVID SANQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID MANAAPLVID CSGKRAVLLD EEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD	160 L S K I D F L D S S G L L S K I D F L D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L T K I D F I D S S G L L T K I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S I D F I D S S G L L S I D F I D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Pm0746 p9211705337 NATL109291 NATL09291 PM70652 P3933_15781 SynRCC307_1330 WH70108484 CW72516 CW72516 CW72516 CW72516 Asr1156 Asr1156 Asr1192 Asr1192 Asr1192 Asr1192 Asr1192 Asr1192 Synf0c7942_2121 Synf0c7942_2121	90 - LROETOPI MHPGGSVPI NKPWGPGPI GSGSI RSIRGFGPI RSIRGFGPI IASROTOPI IIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIII IIIIII	100 SEL - QRL TEL - QRL TE	110 TVSLRGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGFER TVSLRGFER TVSLRGTHEV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTLEV TVSLRGTLEV TVSLRGTLEV TVSLRGTLEV	120 NGCLVFHF KGGCLVFHF RDNCLVFHF KGCLVFYF QNGCLVFYF QNGCLVFYF QNGCLVFYF QOCLVFAF RQGCLVFAF RQGCLVFAF RQGCLVFAF RDNCALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL RDNYALFRL	130 160 L DAYSEK 160 L DAYSEK 161 L DAFSEF 161 L DAFSEF 161 L DAFSEF 161 L DAFSEF 161 L DAFSEF 161 L DAFSEF 160 L DAFSEF 160 L DAFSEF 160 L DAFSEF 160 L DAFSEF 160 L DAFSEF	140 G F ME YVVD VL G F L YVA D VL G F L YY S D VL G F T D F VN D VL G F T D F VN D VF G F T T Y I N E VL G F T T YVE D VL G F T T YVE D VL G F T T YVE D VL G F T T F VL O K VF R K VVG Q YI T F R K VL G K K T F R K VL G K S L T F R K VL G K S L	150 KASKL PAVID KANASSTVD KANASSVVD SSNKL PIVD SSNKL PIVD SSNKL PIVD ASNQLSVVD ASNQLSVVD ASNQLSVVD ASNQLSVVD CSSKRAVLL CSSKRAVLD CSSKRAVLD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHALD	160 L S K I D F L D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S L D F I D S S G L L S L D F I D S S G L L S L I D F I D S S G L L S L I D F I D S S G L L S L I D F I D S S G L L S L I D F I D S S G L L S L I D F I D S S G L L S L I D F I D S S G L L S L I D F I D S S G L L S L I D F I D S S G L L S L I D F I D S S G L L S L I D F V D S S G L L S L I D F V D S S G L L S L I D F V D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L S K I D F I D S S G L S K I D F I D S S G L S K I D F I D S S G L S K I D F I D S S G L
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Pro0746 Pro0746 Pro0746 Pro0746 Pro0745 P	90 - LRGETGPI MHPGGSVPI NKPWGPGPI GSGS RSTRAKPI GTNPI GTNPI IASRGTGPI IASRGTGPI IASRGTGPI CFLARAGTI STNVKEEAI HFNDEEGII HFNDEEGII RFNHEEGII EGVI CCKAEEQVI MTTLEDTAI MTTLEDTAI QSSAKEALI	100 S E L · · Q R L T E L · · Q R L N L N L P E Q · · L N L A E E · · L A L A R E T · · L A L S E P · · L A L	110 TVSL RGGFEQ TVSL RGGFEQ TVSL RGGFEQ TVSL RGGFEQ TVSL RGGFER TVSL RGGFER TVSL RGGFER TVSL RGGFER TVSL RGGFER TVSL RGFEP TVSL RGT REV TVSL RGT LEV		130 T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A F S E H	140 G F ME Y V V D V L G F L T Y V S D V L G F L T Y V S D V L G F L T Y V S D V L G F T T Y I N D V L G F T T Y I N D V L G F T T Y I N D V L G F T T Y I N D V L G F T T Y V E D V L G F T T T Y V E D V L G F T T T Y V E D V L G F T T T T Y V E D V L G F T T T T T T T T T T T T T T T T T T	150 KASKL PAVID KANASSTVID KANTSSLVID SSNKL PIVD SSNKL PIVD SSNQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID CANAAPLVID SSLKPLVID SSLKPLVID COSDKRAVLLD DEGPKHILD DEGPKHILD DEGPKHILD DEGPKHILD DGGPKHAILD ENGPKHAILD EDGPKHAILD EDGPKHAILD	160 L S K I D F L D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L T K I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S I D F I D S S G L L S I D F I D S S G L L S I D F I D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S I D F V D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I E F I D S S G L
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Pro0746 p9211705337 NATL1_09291 PMT0652 PMT0652 P9303_15781 SynRCC307_1330 WH5701_08484 CW72516 Aur4411 n9414703196 Aur4411 n9414703196 Aur4411 syn106728646 Tery_2311 Synpcc 7942_2121 Synpcc 7942_2121 Synpcc 7942_2121	90 - LRGETGPI MHPGGSVPI NKPWGPGPI GSGS RSTRAKPI GTNPI GTNPI IASRGTGPI IASRGTGPI IASRGTGPI CSSKGSRTI CFLARAGTI STNVKEEAI STNVKEEAI HFNDEEGII HFNDEEGII RFNHEEGII RFNHEEGII RFNHEEGII CSSKGAU MTTLEDTAI OCKAEEGVI MTTLEDTAI QSSAKEALI EIDKRRKAI	100 S E L	110 TVSL RGGF EQ TVSL RGGF EQ TVSL RGGF EQ TVSL RGGF EQ TVSL RGGF ER TVSL RGGF ER TVSL RGG SKQ TVSL RGG SKQ TVSL RGG ER TVSL RGG ER TVSL RGG ER TVSL RGT REV TVSL RGT REV TVSL RGT REV TVSL RGT REV TVSL RGT REV TVSL RGT REV TVSL RGT EV TVSL RGT EV TVSL RGT EV TVSL RGT EV TVSL RGT EV		130 T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A Y S E H T G Q L D A F S E F T G L L D A F S E F T G L L D A F S E F T G L L D A F S E F T G L L D A F S E F T G L L D A F S E F T G Q L D A F S E F T G	140 G F ME YVVD VL G F L YVSD VL G F L YVSD VL G F L YVSD VL G F T D F VND VF G F T D F VND VF G F T T YVED VL G F T T T T T T T T T T T T T T T T T T	150 KASKL PAVID KANASSTVID KANTSSLVID QTNKSAVVVD SSNKL PIVD SSNQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID CASAPLVID PSOLKAVLLD EGPYNIILV DEGPKHIILD DEGPKHIILD DEGPKHIILD EGPKHILD EGPKHILD EGPKHILD EGPKHILD EGPKHILD EGPKHILD EGPKHILD EGPKHILD EGSPKHILD EGSPKHILD EGSPKHILD	160 L S K I D F L D S S G L L S K I D F MD S S G L L S K I D F MD S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S Q I D F V D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L L S K I D F I D S S G L
B SynW1027 WH7805_02012 RS9916_28814 RS9917_11490 Pro0746 p9211?05337 NATL_109291 PMT0652 PMT0652 PMT0652 PMT0652 PMT0652 SynC600 Aurat411 n9414?03196 Aurat411 n9414?03196 Aurat411 n9414?03196 Aurat411 synC600 Synf000 -	90 - L R G E T G P I MH P G S V P I NK P W G P G P I G S G S I R S T R R A K P I G T N P I I A S R G T G P I	100 S E L	110 TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFEQ TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGGFER TVSLRGTHEV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV TVSLRGTREV		130 T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G Q L D A Y S E K T G L L D A F S E F T G L L D A F S E F T G L L D A F S E F T G Q L D A F S E F T G	140 G F M YVV D VL G F L YVV D VL G F L YVS D VL G F L YVS D VL G F T T Y N D VL G F T T Y N D VL G F T T Y VE D VL G F T T YVE D VL G F T T T T T T T T T T T T T T T T T T	150 KASKLPAVID KASKSTVID KANTSSLVID QTNKSAVVVD SSNKLPIVID SANQLSVVID ASNQLSVVID ASNQLSVVID ASNQLSVVID MANAAPLVID BSOKRAVLLD EEGPYNIILV DEGPKHIILD DEGPKHIILD DEGPKHIILD DEGPKHIILD EEGSPKNILD EEGSPNIILD EEGSPNILD	160 L S K I D F L D S S G L L S K I D F MD S S G L L S K I D F MD S S G L L S K I D F MD S S G L L S K I D F I D S S G L L T K I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S N I D F I D S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S I D F MD S S G L L S K M M M M M M M M M M M M M M M M M M

FIG. 1. Alignment of sequences of cyanobacterial orthologs of Asr1156. The alignment was produced as shown later in Fig. 6. To conserve space, only part of the alignment is shown. The line for Asr1156 is highlighted in pink. (A) Alignment of nominal protein sequences, according to annotated boundaries. (B) Alignment of protein sequences plus upstream potential amino acids, up to the proximal stop codon.

alignment shown in Fig. 1A. It seems very peculiar that many proteins from organisms you know to be quite distant from each other have regions of very high similarity while other proteins appear truncated.

Suppose it occurs to you that perhaps the annotation is wrong, and what is purported to be the start codon for these short open reading frames are really within the gene. An interesting thought, but one that is very difficult to act on. But suppose further, that you are somehow able to gather

the upstream sequences from all these genes, virtually translate them, and realign them, giving Fig. 1B. There is overwhelming sequence similarity going backwards from the nominal beginning of proteins up to a conserved isoleucine at position 95 in the alignment. Beyond that point, the sequences become dissimilar. It is evident that indeed the start codons must be wrong, and not only for Asr1156 but for every gene in the list! No other explanation can account for such sequence similarity in genes that diverged well over a billion years ago. But why did the gene callers do such an abysmal job? Answer: In most cases there is no conventional start codon between the conserved isoleucine and the next upstream stop codon. The beginnings of the genes are determined by something apart from one of the conventional start codons.

We have described a scenario where the end result is at odds with conventional wisdom, a result ordinarily hidden by the model of molecular biology built into our preexisting tools. At many different points in the journey, most would have filed away the curious circumstances because following up would require abilities that practically speaking we do not have. Without the means to follow up an observation as it is made, chance discovery is lost.

Choices in the face of overwhelming information

Biologists seem to be faced with a list of unpalatable choices:

1. Ignore the wealth of information now available

This choice has the advantage of rooting one's research in tools one understands, thereby making more certain the connection between phenomena and conclusions. A principled approach, perhaps, but how frustrating!

2. Muddle through as best one can with a limited number of tools

This strategy reduces the chance of attaining insights that require a broad understanding of a phenomenon, insights prompted by one of those annoying but precious results that wander outside the confines of our fixed tools.

3. Divide responsibilities amongst those with different expertise

Let the biologists think biological thoughts while working with computer types who think computational ones. This insidiously attractive option relies on the assumption that experimental results are produced by black boxes whose workings are of no scientific interest. We know this to be manifestly false in the case of physical experiments, whose results often depend markedly on exactly how the question is posed. Why should the experimental procedure be less critical for computational experiments? And if we give over the computational experiment to the programmer, then the unavoidable false turns and nonconformant intermediate results we are accustomed to when we work with our own hands will be seen only by someone who does not grasp the biological implications and who does not recognize how a chance observation may attack some basic assumption. By dividing responsibilities, we make rapid gains in predictable ways but slow the pace of insights that would radically change how we look at the world.

4. Entice biologists into learning computer programming

There has been ample incentive over the past 40 years for biologists to learn computer programming. They have voted overwhelmingly not to do so.

One might imagine an additional choice, one that does not yet exist but should: that computer programming become as accessible to typical researchers as word processing, so that anyone can use it to manipulate bioinformation in creative ways with minimal activation energy.

TECO (1970's)

#EBnello.cS\$ Upen file for read/write with backup
#PS
Read in the first page
Silello(serts\$ Search for "Hello" and print the line
printf("Hello worldf\n");
Deloadbye/suit\$ Delete "Hello", insert "Boodbye", and print the line
printf("Boodbye worldf\n");
The updated line
eXS\$

WordStar (1980's)

H:INTRO PAGE 1 LINE 9 COL 11 INSERT ON
< < < MAIN MENU >>>
Cursor MovementDelete- Miscellaneous- _Other_ Menus-
_S char_left _D char_right [_G_ char_lI IabB Reform [(from Main only)
_A word left _F word right DEL chr If V_INSERT ON/OFF _ [_J Help _ K Block
"E line up "X line down [] word rt[E ind/Repice again[0 Uurck "P Print
Scrolling Y line kElukn End paragraph U Unscreen
2 line down w line up N Insert a KElukh
c screen up a screen down i o stop a command i
1. Introducing WordStar
WordStar is highly flexible and very visible. Watch the
screens as you give commands, and information in various
parts of the screen will guide you. You won't see all the
information all the time, but it will be there when you need
it.
RHEKE AOD HKE
The course WandStan manus are usur greatest aids. They are
like signassis at the top of using screens choing your where
und area
THE P 2010 AND 3SET TO ASET RM STINDING ASTRESS ASERDED STRESS AS A ST

Word (2000's)



Figure 2. Evolution of word processors.

The analogy of word processing is informative (3). Just 35 years ago word processing was confined to those who worked with computers professionally. Originally commercial word processing programs were designed for

Fortran (1950's)

		FORTRAN STATEMENT	
NUMBER		1	
c		PROGRAM FOR FINDING THE LARGEST VALUE	
c !	х	ATTAINED BY A SET OF NUMBERS	
		DIMENSION A(999)	
		FREQUENCY 30(2,1,10), 5(100)	
_i		READ 1, N, (A(1), 1-1,N)	
1		FORMAT (13/(12F6.2))	
		BIGA = A(1)	
5		DO 20 I = 2,N	
30		IF (BIGA-A(I)) 10,20,20	
1 10		BIGA = A(I)	
20	_	CONTINUE	
	_	PRINT 2, N, BIGA	
2		FORMAT (22H1THE LARGEST OF THESE 13, 12H NUMBERS 15 F7,2)	
		STOP 77777	

Algol (1960's)

```
procedure Absmax(a) Size:(n, m)
   Result:(y) Subscripts:(i, k);
   value n, m; array a; integer n, m, i, k;
   real y;
begin integer p, q;
   y := 0; i := k := 1;
    for p:=1 step 1 until n do
   for q:=1 step 1 until m do
        if abs(a[p, q]) > y then
            begin y := abs(a[p, q]);
            i := p; k := q
            end
end Absmax
```

Java (2000's)

```
public void insertionSort()
  int in. out;
  for(out=1; out<nElems; out++)</pre>
   long temp = a[out];
   in = out;
   while(in>0 && a[in-1] >= temp)
     a[in] = a[in-1];
     --in;
   a[in] = temp;
  }
```

Figure 3. Evolution of computer languages.

specialist pools within an automated office. This vision of the world did not last for long, as software companies discovered the more lucrative market of individual users. As a result, word processors evolved rapidly (Fig. 2) from clumsy line editors, through full screen editors driven by numerous key combinations, to complex programs that could handle a bewildering number of demands from users, many of whom would consider themselves noncomputational.

Compare this progression of word processors with the succession of programming languages from the 1950's until today (Fig. 3). Computer languages have become far more powerful during that period, but not significantly more usable by those who would consider themselves noncomputational. It is no easier today than it was 40 years ago for a novice to learn how to program a computer.

Is it truly more difficult to encompass all the operations necessary to program a computer than it is to encompass all the operations necessary to run a full-featured word-processor?

We have taken the lessons learned by word processor developers and applied them to an environment that enables biologists to access and manipulate bioinformation in ways limited only by their imaginations. The result is an integrated knowledge and programming environment that

Table 1. Current BioBIKEs^a

CyanoBIKE: Cyanobacteria (37 genomes) ParaBIKE: Eukaryotic parasites (5 genomes) PhotoBIKE: Photosynthetic bacteria (63 genomes) StreptoBIKE: Streptococcus (27 genomes) ViroBIKE: Viruses (1797 genomes and 16 metagenomes) ^aAll instances are available through http://ixion.csbc.vcu.edu/biobike

dispenses as much as possible with the arbitrary syntactical contrivances – the semicolons and parentheses – that so often dissuade biologists from learning the tools they need to make creative use of the knowledge available to them.

BioBIKE Knowledge Bases

We have defined an instance of BioBIKE (**B**iological Integrated Knowledge Environments) as a body of all available information of specific use to a coherent research community. It may seem retrograde to define packets of knowledge in this way, given the trend towards increasingly distributed data linked through the web (4). Doing so, however, brings significant benefits: (a) It is easier to find the data you want, (b) Data can be housed on a single server, greatly reducing access time and permitting calculations that would not otherwise be practical. Users of BioBIKEs have access to two levels of information: rapid access to community-specific knowledge and normal access to outside databases.

At the time of writing there have been five BioBIKEs established (Table 1), all freely available through the web. Each BioBIKE contains genomic sequences, annotation, metabolic function and links to related pathways, and orthologies amongst encoded proteins. In addition, experimental data (e.g. microarrays and results from proteomic experiments) may be available. Similarity scores for all proteins in the database are predetermined, so results of queries regarding protein similarity are available virtually instantaneously. Orthologs (based on bidirectional best hit, per Blast (5) at a threshold of $E = 10^{-6}$) are also precomputed, making it possible to find orthologs common to any subset of organisms in the database within seconds. It is a simple matter, for example, to find in CyanoBIKE all proteins common to, say, nitrogenfixing cyanobacteria that are not found in those that don't fix nitrogen. Protein similarities and orthologs amongst proteins of organisms outside the database are also available, more slowly, through the NCBI implementation of Blast (5) and through KEGG (6). Orthologs may also be determined, more slowly, through any definition provided by the user.

Users are thus spared what is sometimes the grueling task of locating data and converting it to the formats required by different software application.

BioBIKE Programming Environment

The goal of the BioBIKE programming environment is to free the biologist from the burden of arbitrary syntax to take on what should be the more difficult task of formulating biologically meaningful questions. As much as possible, questions that are easy to conceive are also easy to ask. This requires that the language is cognizant of many concepts of molecular biology, such as

HELP! FILE EDIT EXIT PROBLEM ALL ARITHMETIC DEFINITION FLOW-LOGIC	GENES-PROTEINS GENO	ME
INPUT-OUTPUT LISTS-TABLES OTHER-COMMANDS STRINGS-SEQUENCES DATA VARIABLE	DESCRIPTION-ANALYSIS	
	GENE-NEIGHBORHOOD ►	
	GENE-PROTEIN-TYPES 🕨	
	TRANSLATION	\leq
	DESCRIPTION-OF	
	DISPLAY-SEQUENCE-OF	
ORTHOLOG-OF	GENE-OF	
gene-or-protein Options	KEGG-SEQUENCE-OF	
	ORTHOLOG-OF	
	PROTEIN-OF	
	SEQUENCE-OF	

Fig. 4: BioBIKE palette and workspace.

"codon" and "ortholog". The environment adopts conventions common to word processors and web applications, e.g. menus and cut/paste operations.

The fundamental unit of the language is the function, represented as a box. Functions may be selected from one of several menus available from a function palette (Figure 4). The function specifies what input it requires of the user, and the user may supply the information by selecting the entry box and then typing it in or choosing an item from a data menu. The data menu provides (amongst other things) all available organisms and several convenient subsets (Figure 5).

The environment adheres to several design principles:

Intelligibility

The goal of the language is that the meaning of a function should be immediately intelligible to a molecular biologist with no prior experience with BioBIKE.

Limited vocabulary

To reduce the weight of language on the user, a single function may be employed for many different purposes, as directed by options chosen from function-specific menus. For example, the function SEQUENCE-SIMILAR-TO may be used to blast one protein sequence against another protein, against all proteins of one or more organisms in the database, or against all proteins in GenBank. The same function is used to perform nucleotide comparisons or comparisons of translated nucleotides against a protein database, and so forth. The function is also used to find sequences differing from a reference by a given number of mismatches.

HELP! FILE EDIT EXIT PROBLEM ALL ARITHMETIC DEFINITION	ON FLOW-LOGIC GENES-PROTE	NS GENOME
INPUT-OUTPUT LISTS-TABLES OTHER-COMMANDS STRINGS-SEQUENCES	DATA VARIABLES	
	Organism subsets 🔹 🕨	
	Marine unicellular cyano 🛛 🕨	acaryochloris_marina_mbic11017
	Terr. and limnetic unicellular cyano 🕨	crocosphaera_watsonii_wh8501 🔹 🕨
I Workspace prior to browser refresh restored	Filamentous cyano 🛛 🕨	cyanothece_ccy0110
	N fixing cyano 🛛 🕨	prochlorococcus_marinus_as9601 🛛 🕨
ORTHOLOG-OF all4312 IN > Options	microarrays 🕨 🕨	prochlorococcus_marinus_med4
	kegg-pathways 🕨 🕨	prochlorococcus_marinus_mit9211 🛛 🕨
	New data box	prochlorococcus_marinus_mit9301 🛛 🕨
	TRUE	prochlorococcus_marinus_mit9303
	FALSE	prochlorococcus_marinus_mit9312
	GENBANK	prochlorococcus_marinus_mit9313

Fig. 5: Choosing organisms from DATA menu.

ALIGNMENT-OF SEQUENCE-OF ORTHOLOG-OF P-asr1156 Options FROM -100 NO-STOPS LABELED options		ons 🛛
---	--	-------

Fig. 6. Example of nested functions. The nested function makes an alignment of the sequences of all orthologs of the protein Asr1156, starting as many as 100 amino acids before the nominal beginning of the protein but going backwards only up to the first stop codon. The sequences are labeled with the name of the protein and aligned, using Clustal (7) and visualized using JalView (11).

Computability of results

A function may display results in a format designed for ready comprehension, but it will also return the results in a form available for further computation. For example, the results of a Blast performed through SEQUENCE-SIMILAR-TO can be fed directly into a function that performs alignments.

Nesting

Function boxes may serve as input for other function boxes, creating logical phrases (Fig. 6).

<u>Reusability</u>

Functions may be executed repeatedly, perhaps after modifying the contents of an input box. In Fig. 6, the user may re-execute the nested function after replacing the original protein with another.

Multiple arguments

Most BioBIKE functions that make sense to work on both individual items and lists of items do in fact work on both, returning a list of results when provided with a list of arguments. In Fig. 5, for example, the object of the IN option could be a single organism or a list of organisms. This feature greatly reduces the need for loops, the bane of beginning students of computer languages.

Single interface to multiple applications

BioBIKE provides access to several programs that are commonly used: Blast (ref 5; for sequence searches), Clustal (ref 7; for multiple sequence alignments), Meme (ref 8; for motif discovery), RNAz (ref 9; for discovery of conserved RNA sequences), and Phylip (ref 10; for construction of phylogenetic trees). Fig. 7 shows an example of the general interface through BioBIKE function boxes, where the specific behavior of the function is controlled by options chosen from a menu. Figuring out the requirements of free-standing Phylip would be a much more arduous task.



Fig. 7: BioBIKE access to Phylip

Integration of different data types

The examples thus far have focused on genomic information, but BioBIKE a great deal else of interest to a research community. Fig. 8 illustrates how the metabolic knowledge built into BioBIKE, interaction with KEGG, and microarray data can work together to allow a user to find expression values for a specific class of genes. The community can stock the site with microarray

DEFINE =	GENES-IN-PATHWAY/S	*carbohydrate-metabolism [*] in 56803 options option	× S
RATIO-OF C-genes	in "Hilhara2001" at 3	LABELED Options	

Fig. 8: Integration of metabolic and microarray information. Genes from the cyanobacterium *Synechocystis* PCC 6803 (nicknamed S6803) that encode proteins related to carbohydrate metabolism are gathered into a set called "c-genes". A list is then generated of the expression ratios of these genes in the microarray experiment of Hihara et al (12), using the third experimental condition, six hours after a shift from normal to high light intensity. The ratios are labeled with the names of the genes.

data, using a built-in function that can be adapted to a wide variety of different formats. Any data that can be put in the form of a table can be integrated into the knowledge-base.

Porous workspace divisions

All results and functions are confined to the user's personal workspace. However, users are able to share interesting results and functions by placing them in a common workspace. In this way, useful information can be propagated through the community in a form that maximizes its utility. As a body of knowledge of interest to a coherent research community, a BioBIKE may serve individuals of that community but also may help bring the community together.

Styles of BioBIKE Usage

Those new to BioBIKE often begin by using it as a simple query language (e.g., "What is the sequence of my favorite gene?"). From there, the next step is often to construct series of queries, each one dependent on the result of the previous (e.g., "What are the orthologs of my favorite gene?" "What are the upstream sequences of those orthologs?" "What common sequence motifs are found in those upstream sequences?"). This style of usage is made easy by the ability to copy and paste results into the argument boxes of functions and the ability to nest functions.

Many users continue within this style, but the language makes it possible to go much further, automating processes to the same extent as one can with any general purpose computer language. For example, suppose you find that the genome of Anabaena PCC 7120 (A7120) has only 8 instances of the sequence CCCGGG, using the COUNT-OF function (Fig. 9A). This seems intuitively to be too low, but you're not sure, so you repeat the count using a random DNA sequence the same length and nucleotide composition as A7120, and for statistical purposes, you repeat the simulation 3 times (Fig. 9B). Indeed, you find that the expected number of instances of CCCGGG is not 8 but 570. Is this oligomeric sequence uniquely special, or are there others that are also underrepresented? To answer that question, you decide to count every possible 6-nt oligomer, keeping only those that have fewer than 10 counts (Fig. 9C), and find that there are only 12 such hexamers (Fig. 9D). Sorting them leads to the remarkable finding that the 12 can all be fit into one of three restriction enzyme recognition sequences: AvaI (CyCGrG) from Anabaena PCC 7120 itself, AocII (G[AGT]GC[ACT]C) from Anabaena CCAP 1403/9, and NspI (rCATGy) from Nostoc ATCC29411 (a related cyanobacterium). Needless to say, it is highly suspicious that the 12 rarest hexamers (out of $4^6 = 4096$) are recognition sites for enzymes from a close set of relatives!



Fig. 9: Automated re-execution of functions

It is very common to run across a finding that seems intuitively peculiar. Usually these findings remain mere curiosities owing to the difficulty of following up the observations with meaningful tests. Chance discoveries from genomic information, like the one just described, are made possible by the ease with which computational tests can be implemented in BioBIKE.

Whenever a user has devised a set of operations that might be of general utility, the operations can be given a name and packaged into a function, accessible from a menu no differently from any other BioBIKE function. In this way, complicated operations can be broken up into logical chunks, thought of as named items, and used as distinct functions. Complicated chunks can also be collapsed visually into single boxes, making it easier to grasp the larger picture.

General Applicability

BioBIKE is the first attempt we know of to engage biologists without computational experience in the creative analysis of mass bioinformation through an accessible language tailored for their needs. Instances of BioBIKE have been used by researchers, university undergraduates, and high school students, very few of whom have had any prior experience with computer programming. There is no doubt that those without experience can become productive in a very short period of time.

As a first attempt, however, BioBIKE suffers from a number of deficiencies, for example, the sometimes crude representation of results, the incompleteness of documentation for the naïve user, and the deficiency of tools to work with microarrays and other mass data. These deficiencies are being addressed, and we hope that others will also build on our efforts. Any user can add a new function and can share it with others. If deemed useful, it can be integrated into the base language with little difficulty. Those who wish to make deeper changes in BioBIKE can

readily do so, as it is open source, freely available through SourceForge (sourceforge.net). Members of different research communities are invited to build new instances devoted to their favorite groups of organisms, either working with us or on their own.

Programming a computer will never be easy, because it is difficult to formulate questions that are both unambiguous and meaningful. However, formulating such questions is what researchers do for a living. This is not the primary obstacle to enabling biologists to program the computer. That obstacle is the high activation energy that needs to be overcome to gain sufficient proficiency in a language to ask any meaningful question. BioBIKE reduces that activation energy to a minimum. If we biologists can overcome that barrier and embrace computation as a basic tool, we can regain our historical connection with the raw data we study and regain the agility necessary to pursue chance results to new biological insights.

REFERENCES

- 1. Berget, S. M., C. Moore, and P. A. Sharp. 1977. Spliced Segments at the 5' Terminus of Adenovirus 2 Late mRNA. Proc. Natl. Acad. Sci. USA 74:3171-3175.
- 2. Chow, L. T., R. E. Gelinas, T. R. Broker, and R. J. Roberts. 1977. An amazing sequence arrangement at the 5'ends of adenovirus 2 messenger RNA. Cell 12:1-12.
- 3. **Haigh, T.** 2006. Remembering the Office of the Future: Word Processing and Office Automation before the Personal Computer. IEEE Ann. History Computing **28:4:**6-31
- 4. Dowell, R. D., R. M. Jokerst, A. Day, S. R. Eddy, L. Stein. 2001. The Distributed Annotation System. BMC BioInformatics 2:7.
- Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25:3389-3402.
- Kanehisa M., S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa. 2006. From genomics to chemical genomics: new developments in KEGG. Nucl. Acids Res. 34:D354-357.
- 7. Thompson, J. D., D. G. Higgins, T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids Res. 22:4673-4680.
- 8. **Bailey, T.L., C. Elkan.** 1995. Unsupervised learning of multiple motifs in biopolymers using EM. machine learning **21:**51-80.
- 9. Washietl, S., I. L. Hofacker, P. F. Stadler. 2005. Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. U.S.A. 102:2454-2459.
- 10. Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
- 11. Clamp, M., J. Cuff, S. M. Searle, G. J. Barton. 2004. The Jalview Java alignment editor. Bioinf. 20:426-427.
- Hihara, Y., A. Kamei, M. Kanehisa, A. Kaplan, M. Ikeuchi. 2001. DNA Microarray Analysis of Cyanobacterial Gene Expression during Acclimation to High Light. Plant Cell 13:793-806.