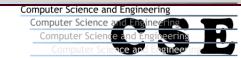
## A Brief Introduction to Bioinformatics



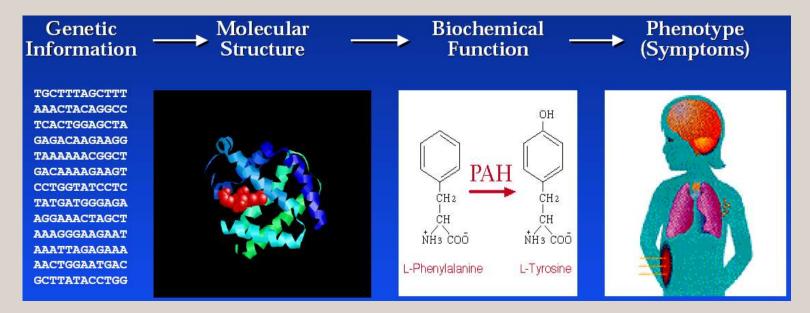
Dan Lopresti Associate Professor Office PL 404B dal9@lehigh.edu





### Motivation

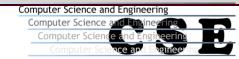
"Biology easily has 500 years of exciting problems to work on." Donald Knuth (Stanford Professor & famous computer scientist)



By developing techniques for analyzing sequence data and related structures, we can attempt to understand genetic nature of diseases.

http://cmgm.stanford.edu/biochem218/





### Bioinformatics

What is bioinformatics?

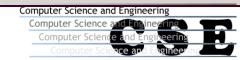
Application of techniques from computer science to problems from biology.

Computer		
	Bioinformatics	
	Biology	

Why is it interesting?

- Important problems.
- Massive quantities of data.
- Desperate need for efficient solutions.
- Success is rewarded.



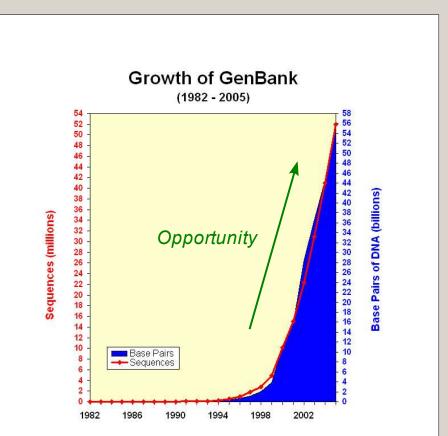


### Data Explosion

Genetic identity of most organisms is encoded in long molecules made up of four basic units, the nucleic acids:

- (1) Adenine,
- (2) Cytosine,
- (3) Guanine,
- (4) *Thymine*.

To first approximation, DNA is language over 4 character alphabet,  $\{A, C, G, T\}$ .



http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html



Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering

### Genomes

### Complete set of chromosomes that determines an organism is known as its *genome*.

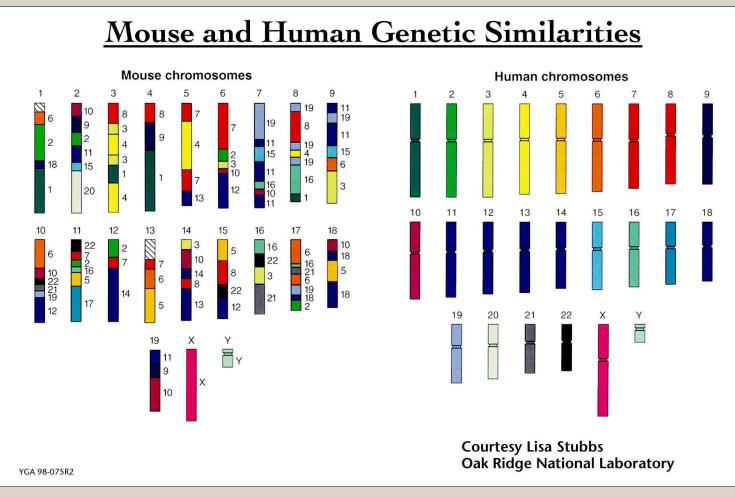
liowit as its genome.	GenBank Release 121.0 -	- December 15, 2000		
	Species	Haploid genome size	Bases	Entries
Us ————	Homo sapiens	3,400,000,000	6,702,881,570	3,918,724
	Mus musculus	3,454,200,000	1,291,602,139	2,456,194
	Drosophila melanogaster	180,000,000	487,561,384	166,554
	Arabidopsis thaliana	100,000,000	242,674,129	181,388
AND A CONTRACTOR	Caenorhabditis elegans	100,000,000	203,544,197	114,553
Poaceae	Tetraodon nigroviridis	350,000,000	165,539,271	188,993
	Oryza sativa	400,000,000	125,948,974	151,411
	Rattus norvegicus	2,900,000,000	106,344,366	218,598
	Bos taurus	3,651,500,000	71,215,626	159,473
Mug mugaulug	Glycine max	1,115,000,000	62,817,102	141,802
Mus musculus	Medicago truncatula	400,000,000	50,991,920	104,535
	Trypanosoma brucei	35,000,000	49,855,996	91,334
	Lycopersicon esculentum	655,000,000	49,415,566	97,112
	Giardia intestinalis	12,000,000	47,639,714	54,328
Zoo movo	Strongylocentrotus purpur	900,000,000	47,590,936	77,532
Zea mays	Entamoeba histolytica		44,522,016	49,938
	Hordeum vulgare	<u>10 - </u> 71	44,489,692	57,779
	Danio rerio	1,900,000,000	40,906,902	83,726
	Zea mays	5,000,000,000	36,885,212	77,506
⊧p://www.cbs.dtu.dk/databases/DOGS/	Saccharomyces cerevisiae	12,067,280	32,779,082	18,361

http://www.nsrl.ttu.edu/tmot1/mus\_musc.htm http://www.oardc.ohio-state.edu/seedid/single.asp?strID=324



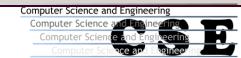


### **Comparative Genomics**



http://www.ornl.gov/sci/techresources/Human\_Genome/graphics/slides/ttmousehuman.shtml



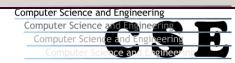


### Algorithms are Central

An *algorithm* is a precisely-specified series of steps to solve a particular problem of interest.

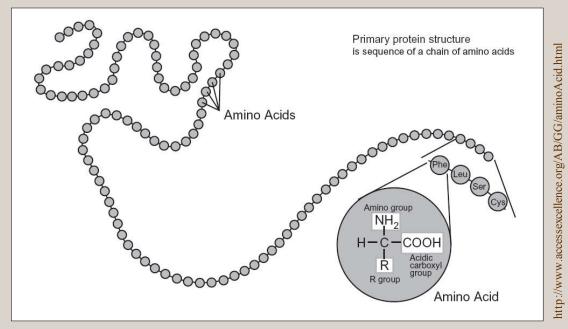
- Develop model(s) for task at hand.
- Study inherent computational complexity:
  - Can task be phrased as an optimization problem?
  - If so, can it be solved efficiently? Speed, memory, etc.
  - If we can't find a good algorithm, can we prove task is "hard"?
  - If known to be hard, is there approximation algorithm (one that works at least some of the time or comes close to optimal)?
- Conduct experimental evaluations (perhaps iterate above steps).





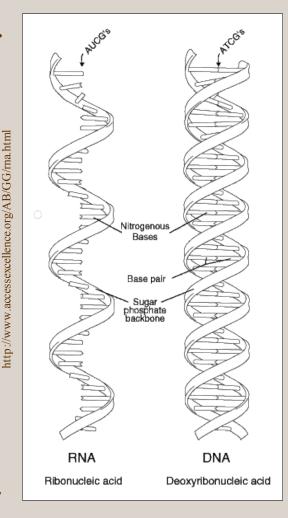
### Sequence Nature of Biology

#### Macromolecules are chains of simpler molecules.



In the case of proteins, these basic building blocks are *amino acids*.

In DNA and RNA, they are nucleotides.



Computer Science and Engineering

Computer Science



### NCBI GenBank

National Center for **Biotechnology Information** (NCBI), which is branch of National Library of Medicine (NLM), which is branch of National Institutes of Health (NIH), maintains *GenBank*, a worldwide repository of genetic sequence data (all publicly available DNA sequences).

/db_xref="GeneID:3098339" /translation="MKNISDLWNQALGQIEKKLSKPSFETWMKSTKAHSLQGDTL: APNFFARDWLESRYLHLIADTIYDLTGEELSIKYVIPONONEEDFMFKSPIKKMSJ PADFPONMLNPKYTFDTFVIGSGNRFAHAASLAVAEAPAKAYNPLFIYGGVGLGK MHAIGHYVIDHNPSAKVVILSEKTNEFINSIRDNKAUDFRNRYNVVULIDD LAGKEQTQEEFFHTFNTLHEETKQIVISSDRPPKEIPTLEDRLRSRFEWGLITDI DLETRIAILRKKAKAEGLDIPNEVMLYIANQIDSNIRELEGALIRVVAYSSLINKI ADLAAEALKDIIPSSKPKVITIKDIQRIVGQQFNIKLEDFKAKKRTKSVAFPRQIJ LSREMTDSSLPKIGEEFGGRDHTTVIHAHEKISKLLSDDEQLQQQIKEIKEQLR" ORIGIN 1 atgaaaaaca tatcggatct ttggaatcaa gctttggggc agatcgaaaa aaaattgagc 61 aagcccagct ttgaaacatg gatgaaatcg acaaaggccc attcattgca gggcgatacg 121 ctgatcatca ccgcaccgaa cgagttcgcc agagatggc ttgaatcaag atacctgcac 181 ctgatcgccg atacggatca tgatctgaca ggagaagaat tgggcattaa atttgtcatt 241 cctcagaatc aaaatgaaga agatttatg ccaaagtcc caatcaaaaa aatgtcgaa 301 gaagaaccgg ctgatttcc gcaaaacatg ctgaatccaa aatatacatt tgatacgttc 361 gttatcggt caggaaacg attcgcca gaggggtcg gactggaaa gactcatta 481 atgcatgga tcgggcacta tgtcatcgat cacaaccat ctgcaaaag ggtttattg 541 tcatctgaga aatttacaa tgagtcat aactcgatc cgtgacaataa agtgtcgat 661 ggaaaagaac agacgcaaga ggaattttc catacgtcc gtgacaataa agtgtcgat 661 ggaaagaac agacgcaag ggaattttc catacgtca cgtgacaataa agtgtcgat 661 ggaaagaac agacgcaag ggaatttat aaccgatcg tgaagaaca 721 aagcagatg tcattccaa tggtcgt cctttaatag acgatatta attttagcc 661 ggaaagaac agacgcaag ggaattttc catacgtta ataccgtcg tgaagaaca 721 aagcagatg tcattccag cgaccgcc ccaaaagag tccaaacgt tgaagaaca 641 agaatgga tttaagaa gaagcaaag gcgaagga ttgaatccc gaagaagac 641 agaatggg tttaagaa gaagcaaag gcgaagga ttgaatccc gaagaagac 641 agaatgga tttaagaa gaagcaaag cagaaggac ttgataccc gaagaagac 641 agaatgga tttaagaaa gaagcaaag cagaaggac ttgataccc gaagaagac 641 agaattgcg ctattctc cttgatcaat aaactaggg agtgattacc 641 agaattgcg ctattctc cttgatcaat aaactagg agtgtgata acc 641 agaattgcg ctattctc cttgatcaat aaactagga agtgtatac 641 agaattgcg ctattctc cttgatcaat aaactagga actggattaac	
<pre>/db_xref="GeneID:3098339" /translation="MKNISDLWNQALGQIEKKLSKPSFETWMKSTKAHSLQGDTL' APNEFARDWLESRYLHLIADTIYDLTGELSIKFVIPQNQNEEDFMPKSPIKKMSJ PADFPQMULNPKYTFDTFVIGSGNRFAHAASLAVAEAPAKAYNPLFIYGGVGLGK' MHAIGHYVIDHNPSAKVVYLSSEFTNEFINSIRDNKAVDFNNRYNVDVLLIDD LAGKEQTQEEFFHTFNTLHEETKQIVISSDRPPKEIPTLEDRLRSRFEWGLITDI' DLETRIAILRKKAKAEGLDIPNEVMLYIANQIDSNIRELEGALIRVVAYSSLINKI ADLAAEALKDIIPSSKPKVITIKDIQRIVGQOFNIKLEDFKAKKRTKSVAFPQIJ LSREMTDSSLPKIGEEFGGRDHTTVIHAHEKISKLLSDDEQLQQQIKEIKEQLR" ORIGIN 1 atgaaaaaca tatcggatt ttggaatcaa gctttggggc agatcgaaaa aaaattgagc 61 aagcccagct ttgaaacatg gatgaaatcg acaaaggccc attcattgca gggcgatacg 121 ctgatcatca ccgcaccgaa cgagttcgcc agagaatggc ttgaatcaag atacctgcac 181 ctgatcgcg atacggatta tgatcgaca ggagaagaat tggcattaa atttgtcatt 241 cctcagaatc aaaatgaaga agattttatg ccaaagtcc caatcaaaaa aatgtcgaaa 301 gaagaaccgg ctgatttcc gcaaacatg ctgaatccca aatatacatt tgatacgttc 361 gttatcggt caggacata tgcaccac gcaggtgct tggaagaa gactcatta 481 atgcatgga tagtacata tgtcaccgat cacaatccat ctgcaaaag ggtttattg 541 tcatctgaga aatttacaa tgagtcatt aactcgatcc gtgacaataa agctgcgat 661 ggaaagaac agacgcaaga ggaattttc catacgttc attataca attttagcc 661 ggaaagaac agacgcaag ggaattttc catacgttc attatacat ttgacgtcg 541 tcatctgaga aatttacaa tgggtcatt aactcgatcc tgagaaatac 661 ggaaagaac agacgcaag ggaattttc catacgtta ataccgctg tgaagaaca 721 aagcagatg tcattccag cgaccgcc ccaaaagaga tccaaacg tgaagaaca 841 agaatgga tttaagaa gaagcaaa gcagaaggac ttgataccc gatgaagac 841 agaattgcga tttaagaa gaagcaaa gcagaaggac ttgataccc gatgaagac 961 atgcttaa ttgccaatc gatcgacca aattacagg ggcattaatc 961 agggttgcg catattctc cttgatcaa aatacagg ggcattaat 961 agggttgcg catattctc cttgatcaa aatacagga ggcattaatc 961 agggttgcg catattctc cttgatcaat aacagcagt ggcattaac 961 agggttgcg catattctc cttgatcaat aaagacatta acgccgatc gatgaagac 961 agggttgcg catattctc cttgatcaat aaagacatta acgccgatc ggcagcaga 961 agggttgcg catattctc cttgatcaat aaagacatta acgccgatc ggcagcaga 961 agggttgcg catattctc cttgatcaat aaagacatta acgccgatc</pre>	
<pre>/db_xref="GeneID:3098339" /translation="MKNISDLWNQALGQIEKKLSKPSFETWMKSTKAHSLQGDTL' APNEFARDWLESRYLHLIADTIYDLTGELSIKFVIPQNQNEEDFMPKSPIKKMSJ PADFPQMULNPKYTFDTFVIGSGNRFAHAASLAVAEAPAKAYNPLFIYGGVGLGK' MHAIGHYVIDHNPSAKVVVLSSEFTNEFINSIRDNKAVDFNNRYNVDVLLIDD LAGKEQTQEEFFHTFNTLHEETKQIVISSDRPPKEIPTLEDRLRSRFEWGLITDI' DLETRIAILRKKAKAEGLDIPNEVMLYIANQIDSNIRELEGALIRVVAYSSLINKI ADLAAEALKDIIPSSKPKVITIKDIQRIVGQOFNIKLEDFKAKKRTKSVAFPRQIJ LSREMTDSSLPKIGEEFGGRDHTTVIHAHEKISKLLSDDEQLQQQIKEIKEQLR" ORIGIN 1 atgaaaaaca tatcggatct ttggaatcaa gctttggggc agatcgaaaa aaaattgagc 61 aagcccagct ttgaaacatg gatgaaatcg acaaaggccc attcattgca gggcgatacg 121 ctgatcatca ccgcaccgaa cgagttcgcc agagaatggc ttgaatcaag atacctgcac 181 ctgatcgcg atacggatca tgatcgaca ggagaagaat tggcattaa atttgtcatt 241 cctcagaatc aaaatgaaga agattttatg ccaaagtctc caatcaaaaa aatgtcgaaa 301 gaagaaccgg ctgattttc gcaaacatg ctgaatccca atatacatt tgatacgttc 361 gttatcggt caggacacg attcgccca gcaggcgctt tggcaggag tgaagcccg 421 gcgaaagct acaatccgt gtttattac gggggagtcg gactggaaa gactcatta 481 atgcatgga taggcata tgtcatcgat cacaatccat ctgcaaaag ggtttattg 541 tcatctgaga aatttacaa tgagtcatt aactcgatcc gtgacaataa agctgcgat 661 ggaaagaac agacgcaag ggaattttc catacgttc attacaat ttgacgtcg 721 agcgatg tcattccag cgacggct ccaaacaag ctgaaaa 721 aagcagatg tcattccag cgacggct ccaaacacg tgaagaaca 721 aagcagatg tcattccag cgacggct ccaaaagag tccaagct tgaagaaca 841 agaattgcga ttttaagaa gaagcaag agcagagga ttggatacc gactggaaca 841 agaattgcg ttttaagaa gaagcaaa gcagaaggac ttgataccc gaagaagc 901 atgctttaa ttgccaatca gaccgacca atatcagg agcggatcg 901 atgctttaa ttgccaata gatcgacca aaatacaat acgccgact gacgaaca 901 agggttgcg catattcttc cttgatcaat aacacacac gccgact gacgaaca 901 agggttgcg catattctc cttgatcaat acgccgacg 901 atgctttaa ttgccaata gaccgacaga agcagaaca 901 agggttgcg catattctc cttgatcaat aaagacatta acgccgact ggctgcaaa 901 agggttgcg catattctc cttgatcaat aaagacatta acgccgact ggctgcaaa 901 agggttgcg catattctc cttgatcaat aaagacatta acgccgact ggctgctgaa</pre>	
APNEFARDWLESRYLHLIADTIYDLTGEELSIKFVIPONONEEDFMPKSPIKKNS PADFPONMLNPKYTFDTFVIGSGNRFAHAASLAVAEAPAKAYMPLFIYGGVGGG MHAIGHYVIDHNPSAKVVYLSSEKFTNEFINSIRDNKAVDFRNRYRNVDVLIDD LAGKEQTQEEFFHTFNTLHEETKQIVISSDRPKEIPTLEDRLRSRFEWGLIDD DLETRIAILRKKAKAEGDIPNEVMLYIANQIDSNIRELEGALIRVVAYSSIINKI ADLAAEALKDIIPSSKPKVITIKDIQRIVGQQFNIKLEDFKAKKRTKSVAFPROJ LSREMTDSSLPKIGEEFGGRDHTTVIHAHEKISKLSDDEQLQQQIKEIKEQLR" ORIGIN 1 atgaaaaca tatcggatt ttggaatca gctttgggc agatcgaaa aaaattgagc 61 aagcccagt ttgaaacag gatgaaatcg acaaaggcc attaattga gggcgatacg 121 ctgatcata ccgcaccga cgagttcgc agagatgg ttgaatcaa attgtcat 181 ctgatggccg atacgatct tggtaatca ggagaagat tgagcataa attgtgaaa 301 gaagaaccg ctgatttc gcaaaacag ctgagatct tggaatcaa attgtgaaa 301 gaagaaccg ctgatttc gcaaacag ctgaggtcg gacttgaaa gattcatg 421 gcgaaagt acaatcgg attgacta tgactgaa gactgaaa agactaata 481 atgcatgga taggacca ttgtacgaa caaaggcc gactggaa gactacta 481 atgcatgga taggacca tgtcatga caaagtc tggaaaaa gactacata 661 ggaaagaac gattacaa tgatgaatt aactggac gactggaa gactacta 661 ggaaagaac gagtacaa tgtcagga caaagtat tggacaata agtggaaa 721 aagcagatg tcattcaa tgagtcat aactgacaa dcgaagaac tgaagaaca 721 aagcagatg tcattcaa gggaatttte caaagaga tccaaact tgaagaaca 721 aagcagatg tcattcaa ggagacg ccaaaggac tgaatccc aattacaat tgaagaaca 721 aagcagatg tcattcaa ggagaacaa gcgaaagga tccaaagag tgaagaaca 721 aagcagatg tcattcaa gagagaaa gaagcaaa gcgaaagga tccaaacc tgaagaaca 721 aagcagatg tcattcaa gagagaa ggaatttte cataggac tgaagaaca 721 aagcagatg ttatagaa gaagcaaa gcgaaagga ttgaacatc gaagaaca 721 aagcagatg tcattcaa gaagcaaa gcgaaagga ttgaaccc gaagagac 731 ttgcgctcc gtttgaag ggaattac actgaacaa accgcg ttgaaacc 731 tgcgatca gtttgaaga gaagcaaa gcagaagga tgaatccc gaagaagc 731 tgcgattga tttaagaa gaagcaaa gcagaagga tgaatccc gaagaagac 731 aggattggg tttaagaa gaagcaaa gcagaagga tgaatccc gaagaagc 731 atgcattaa ttgccaata gacgaaa gaagaaa gcagaagga tgaatccc gaagaagac 731 atgcattga tttaagaa gaagcaaa gcagaagga ttgaatcc gaagaagac 731 atgcattga tttaagaa gaagcaaaa gcagaagag ttgaatcc gaagaagac 731 atgcattga tttaagaa gaagcaaa gca	
PADFPONMLNPKYTFDTFVIGSGNRFAHAASLAVAEAPAKAYNPLFIYGGVGLGK: MHAIGHYVIDHNPSAKVVYLSSEKFTNEFINSIRDNKAVDFRNYRNVDVLLIDD LAGKEQTQEEFFHTFNTLHEETKQIVISSDRPPKEIFTLEDRLRSPFEWGLIDD LAGKEQTQEEFFHTFNTLHEETKQIVISSDRPPKEIFTLEDRLRSPFEWGLIDD DLETRIAILRKKAKAEGLDIPNEVMLYIANDQOFNIKLEDFKAKKRTKSVAFPRQI JLAAFALKDIIPSSKPKVITIKDIQRIVGQOFNIKLEDFKAKKRTKSVAFPRQI LSRENTDSSLPKIGEFGGRDHTTVIHAHEKISKLLSDDEQQQIKEIKEQLR" ORIGIN 1 atgaaaaaca tatcggatt ttggaatca gcttggggc agatcgaaa aaaattgagc 61 aagccagt ttgaacatg gatgaatcg acaaaggcc attaatgca gggggatacg 121 ctgatcatca ccgcaccgaa cgagtcgcc agagactgg ttgaatcaag atacctgcac 181 ctgatcgc atacgatt tgatctgac ggagaagat tgagcataa attgtcgaaa 301 gaagaaccg ctgatttcc gcaaaacatg ctgaatcca aatatacat tgatacgtc 361 gttatcggt caggaaccg attgtcgc agagggtg gattgaaa aagtcgaaa 301 gaagaaccg ctgatttcc gcaaaacatg ctgaatcca ctgcaaagag ggttattg 541 tcatctgaa taggccata tgtcatcgat cacaatccat ctgcaaagag ggttattg 541 tcatctgaa aattgaaa tgtgacgt cttttaatag acgatatca attttagcc 661 ggaaaagaa agattttc catacgtta atacgtgca tgaagaaca 721 aagcagattg tcattccag cgaccgcc ccaaaagag tcccaatca tgtaggaaca 721 aagcagattg tcattccag cgaccgcc ccaaaagag tcccaatca 841 agaattgcg ttttaagaa gaagcaaa gcagaagga ttgataccc gaagaaca 841 agaattgcg ttttaagaa gaagcaaa gcagaaggc ttgataccc gaagaaca 901 atgctttaa ttgccaata gaccgaca aatccag cgacggac ttgatacc 901 agggttgcg catattctc cttgatcaa aaagacatt accgcacc gaagagc 901 atgctttaa ttgccaata gatcgaca aaagagaaca acgcgacgt ttgatacc 901 agggttgcg catattctc cttgatcaa aaagacatt accgcacc gacgagac 901 agggttgcg catattctc cttgatcaa aaagacatta accgcacc ggccgaccacca 901 agggttgcg catattctc cttgatcaat aaagacatta accgcacct ggccgacacc 901 agggttgcg catattctc cttgatcaat aaagacatta accgcacct ggccgcaccacca 901 agggttgcg catattctc cttgatcaat aaagacatta accgcacct ggccgcacca 901 agggttgcg catattctc cttgatcaat aaagacatta accgcacct ggccgcaccaccacct ggccgaccaccaccaccaccaccaccaccaccaccaccac	TEE
<ul> <li>MHAIGHYVIDHNPSAKVVYLSSEKFTNEFINSIRDNKAVDFRNRYRNVDVLLIDD: LAGKEQTQEEFFHTFNTLHEETKQIVISSDRPPKEIPTLEDRLRSRFEWGLITDI DLETRIAILRKKAKAEGLDIPNEVMLYIANQIDSNIRELEGALIRVVAYSSLINKI ADLAAEALKDIIPSSKPKVITIKDIQRIVGQQFNIKLEDFKAKKRTKSVAFPRQID DLETRIAILRKKAKAEGLDIPNEVMLYIANQIDSNIRELEGALIRVVAYSSLINKI ADLAAEALKDIIPSSKPKVITIKDIQRIVGQQFNIKLEDFKAKKRTKSVAFPRQID USREMTDSSLPKIGEEFGGRDHTTVIHAHEKISKLESDDEQLQQQIKEIKEQLR"</li> <li>Argaaaaaa</li> <li>aagcccagci ttgaacaag gatgaagac gactagci ttgaatcaag adaattgagc 121 ctgatcatca ccgcaccgaa cgagttgcc agaagactgc ttgaatcaag attactt 241 cctcagaac aaatgaag agattttag ccaaagtcc caatcaaaa aatgtcgaa 301 gaagaaccgg ctgatttcc gcaaaacatg ctgaatcca aatatacatt tgatacgttc 361 gttatcggt caggacac gttatttag ggggagagg gactggaag gactgattag 421 gcgaaget acaatccgc gtttattac gggggagtg gactgaaga gactcatta 481 atgcatgga tcgggcacta tgtcatcgat cacaatccat ctgcaaaag ggttattgg 541 tcatctgaa aattacaaa tgtgacgt ctttaatag acgatatca attttagcc 661 ggaaagaac ggacagag ggaattttc catacgtta accagatca tgaagaacag 721 aagcagattg tcattccag cgaccgcet ccaaaagag tcccaata agtgtcgat 661 tttegcaac gttttgaag gggattgat actcgaatca cccaaccet tgagaacaa 721 aagcagattg tcatttccag cgaccgcet ccaaaagag tcccaag tcggaacaa 841 agaattgga tttaagaa gaagcaaag gcagaaggc ttgatacce gatgaaca 841 agaattgga tttgaaga gatcaa aataccagg agtgtgat cattacag tcggaacac 901 atgctttaa ttgccaata gatgacaa gcagaagag gactatac 901 atgctttaa</li> </ul>	
Address and a set of the set of t	
DLETRIAILRKKAKAEGLDIPNEVMLYIANQIDSNIRELEGALIRVVAYSSLINK ADLAAEALKDIIPSSKPKVITIKDIQRIVGQOFNIKLEDFKAKKRTKSVAFPRQI LSREMTDSSLPKIGEFGGRDHTTVIHAHEKISKLLSDDEQLQQOIKEIKEQLR DRIGIN 1 atgaaaaaa tatgggatt tgggaatag acaaagggc attatag gggggaagg 121 ctgatcaca ccgcaccga cggttcgc agagacgg ttgaatcaa attgac 181 ctgatcgcc aaagtgaa agatttat ccaaagtcc caatcaaa aatgcgaa 181 ctgatcgcc aaagtgaag agatttat ccaaagtcc caatcaaa aatgcgaa 181 ctgatcgt caggaaacg attggcca ggggaggag aaattat tgatcatt 241 cctcagaac caatgaag agatttat cgaaagtcc aattacaat tgatcgaa 301 gaagaaccg ctgatttc gcaaaasag cgaggagg gatt 361 gtatcggt caggaagca ttggccaa tgcaaata aggcgcg 421 gcgaaagt acaatccgt gttattac gggggagtg gactggaa ag 481 atgcatgga tcggcaata tgcacgaa caaactac tgcaaaaa agtcgaa 661 ggaaagaa agattacaa tgtgagtt ctttaatag acgatataa atttagc 661 ggaaagaa agacgaag ggaattttc catacgtta acccaacca tgaaaaa 721 aagcagatg tcattccag cgaccgcc ccaaagaga tcccaacc tgaaaaaa 721 aagcagatg tcattccag cgaccgcc ccaaagaga tcccaacc tgaaacaa 721 aagcagatg tcattccag cgaccgcc ccaaagaga tcccaacc tgaaacaa 721 aagcagatg tcattccag cgaccgcc ccaaagaga tcccaacc tgaaacaa 721 aagcagatg ttatacaa gaagcaaa gcagaagga ttgatccc gaagaacaa 721 aagcagatg ttatagaa gaagcaaa gcagaagga ttgatccc gaagaaca 721 aagcagatg ttatagaa gaagcaaa gcagaagga ttgatccc gaagaaca 721 aagcagatg ttatagaa gaagcaaa gcagaagga ttgatccc gaagaaca 721 aagcagatg ttatagaa gaagcaaa gcagaagga ttgatacc caaccc tgaaaacaa 721 aagcagatg ttatagaa gaagcaaa gcagaagga ttgatccc gaagaaca 721 aagcagatg ttatagaa gaagcaaa gcagaagga ttgatccc gaagaacc 731 ttgcgctcc gtttgaag ggaattac actgacacc acccc tcgaaaca 731 tgcgttaa ttgccaata gaagaaa gaagaaga tgaaacc 731 aggattgg ttataagaa gaagcaaa gcagaagga ttgatcc gaagaacc 731 aggattgg ttataagaa gaagcaaa gcagaagga accc gaagaga gcattaac 731 aggattgg ttatagaa gaagcaaa gcagaagaga ttgaaccc gaagagc 731 aggattgg ttatagaa gaagcaaa gcagaagaga ttgaatccc gaagaaca 733 aggattgaa ttgcaaaa gaagaaa gaagaaaa gcagaagaa caaccg ttgaaaca 734 agaattgg ttatccaa caaccaac aaccaaccaa tcaaccaac ttataag	1
ADLAAEALKDIIPSSKPKVITIKDIQRIVGQOFNIKLEDFKAKKRTKSVAFPROI LSRENTDSSLPKIGEEFGGRDHTTVIHAHEKISKLSDDEQLQQQIKEIKEQLR DRIGIN 1 atgaaaaca ttggatt ttggaataa gattgaagg agattgga agattggg agagaga ttgatcaag agattgaa 11 atgatacga taagatta tgatcgaa gagagaga tgaataag attggaat 121 ctgatcaaa aaatgaag agatttag ccaaagtot tgaacaag attgtgat 131 ctgatcgg ataggatga agatttag ccaaagtot tgacataa attgtgat 131 ctgatcgg ataggatga agatttag ccaaagtot tgacataa attgtgat 131 gaagaacgg ctgatttc gcaaaacatg ctgaatcca aatacaat tgatacgta 131 gaagaacgg ctgatttc gcaaaacatg ctgaatcca aatacaat tgatacgta 131 gaagaacgg ctgatttc gcaaaacatg ctgaatcca aatacaat tgatacgta 131 gaagaacgg ctgatttc gcaaaacatg ctgaatcca attgcaag gactgat 131 tcaccgta tgggacat tgtcacca gagggtg gactgat 131 tcaccgaa tgggacaa tgttgacgt ctttaatag acattgaag ggttgat 131 tcaccgaa agacgaag ggaattttc catacgatc gtgaatcaa actgatca 132 ggaaagaac agacgcaag ggaattttc catacgatc gtgaatcaa 133 agacagatg tcattccag ggggtgat actgaatca tgtagaaca 134 agacatgga ttttaagaa gaagcaaa gcagagga ttgatacg tgaagacg 131 ttgggctcc gtttgaag ggaatgaa gcaacaag tcgaaaca tgtagaaca 132 agagattga ttttaagaa gaagcaaa gcagaagga tgaatacc tgaaacac 133 agaattgga ttttaagaa gaagcaaa gcagaagga tgaatca tgaagacg 134 agaattgga ttttaagaa gaagcaaa gcagaagga tgatacc gaagaaga 134 agaattgga ttttaagaa gaagcaaa gaagaaga tgagaaca gaagaaga tgatacc 134 agaattgga tttaagaa gaagcaaa gcagaaga tgatacca gacggaaga tgatacc 135 aggattga tttaagaa gaagcaaa gcagaaga tgatacca gacgaaga tgatacc 136 aggattga tttaagaa gaagcaaa gcagaaga tgatacca gacgaaga tgatacc 137 agaacgatg tttaagaa gaagcaaa gaagaaga tgatacca gaagaaca 138 agaattga tttaagaa gaagcaaa gcagaaga tgatacca gaagaaca 139 aggttgat tttaagaa gaagcaaa gcagaaga tgatacca gaagaaca tgatagaac 139 aggttgat tttaagaa gaagcaaa gcagaaga tgatacca gaagaaca fgagaaca 139 aggttgat tttaagaa gaagcaaa gcagaaga tgatacca gaagaaca tgaagcaa gaagaaca gaagaaca fgagaaca fgagaaca fgagaaca fgagaata fgagaaca fgagaaca fgagaaca fgagaaca fgagaaca fgagaaca fgagaata fgagaaca fgagaaca fgagaaca fgagaaca fgagaata f	
LSRENTDSSLPKIGEEFGGRDHTTVIHAHEKISKLLSDDEQLQQQIKEIKEQLR"DRIGIN1 atgaaaacatatcggattttggaatcagctttgggcagatcgaaagaaatgagc11 atgaaaacatatcggattttggaatcagctttgggcatactgcaaggggcgatacg121 ctgatcatcaccgacacgacggatcgacgagaatgaatactgcaaatactgcaa181 ctgatcgccatacgatctatgacactgacgagaagaattgacattaaatatgcgaa201 gaagaaccgctgattccgcaaaacatgctgaatccaaatatgctgcaatagccccg301 gaagaaccgctgatccccgttatcggtcatacaaaaatatgctgaaggttgaaacggttatcgtt301 gaagaaccgctgggaaacggttatcggtcatacaaaaatgtcgaaggttgaatgaagagcgtgggcttggaaggttatcggt311 tcatcggtcaggaacgagttatcggtgttatcggtgcttatcggtgcttatcggtgaagcccgggcttatcggt321 gcgaaagtcagggcactatgtcatcgattgtcatcgatgcgttacgatggtttatcggcttatcggt321 tcatcgaaggagcaagaggattttatacactgcacctgaaaacagctgcaagatgttatgcgt341 tcatcgaaagaacgcaagaggatttttcatacgataittagccittagaaacaittagcatgg321 tgggctccgttttgaagggattgatcacatcgataittagaacgittagaacgittagaacg341 tcatcgaatgaacgcaagaggattgatcacatcgataittagaacgittagaacgittagaacg341 tcatcggatttcattccagggacggatgatactgatacaittagaacgitgaaacgitgaaacg341 tca	
DRIGIN 1 atgaaaaaca tatcggatci tiggaatcaa gettigggge agategaaaa aaaatigage 61 aageecaget tigaacatg gatgaaateg acaaaggeec atteatigea ggeegataeg 121 etgategee acaegateg etgategee agagaetgge tigaateaag ataeetgeae 181 etgategee aaaatgaaga agattitatg ecaaagtee caateaaaa aatgeegaaa 301 gaagaacegg etgattite geaaaacatg etgaateeaa aatateatt tigataegte 361 gitateggit eaggaaaeeg attegeecae geagegiett tiggeaggae tigaageeeg 421 gegaaageet acaateege gitattitee geaaaaete eggeggeet tiggaateaa aatgeegaa 301 gaagaacegg etgattitee geaaaaeatg etgaateeaa aatateett tigataegite 361 gitateggit eaggaaaeeg attegeecae geagegiett tiggeaggae tigaageeeg 421 gegaaageet acaateeget gittattae gggggagteg gaettiggaaa gaeteaetta 481 atgeatgega tegggeaeta tigtategat caeaateeat etgaeaaag ggittattig 541 teatetgaa aattaeaa tigagteet aateegee gittaateg attegeega 601 titegeaate getategaaa tigtigaegit etittaatag aegatatea attittagee 661 ggaaaagaae agaegeaaga ggaattite cataeagta aceaatee tigaagaaeaa 721 aageagattg teatteeag eggeate eeaaaaggaggae tigaateee tiggaaaeaa 841 agaattgega tittaagaa gaageaaa geegaaggae tigaateee gaagaagge 901 atgettata tigecaatea gategeaea aateeegg agetggaag ggeattaate 961 agggitigee catateete etigateaa aaageeata aaageegte gaegaagge	.MY
1 atgaaaaaca tatcggatct ttggaatcaa gctttggggc agatcgaaa aaaattgagc 61 aagccagct ttgaaacatg gatgaaatcg acaaaggccc attattgca gggcgatacg 121 ctgatcatca ccgcaccgaa cgagttggc agagaatggc ttgaatcaag atactgcac 181 ctgatcgcg atacgatcta tgatcgaca ggagaagaat tggcattaa attgtcatt 241 cctcagaatc aaaatgaaga agatttatg ccaaagtccc caatcaaaaa aatgtcgaaa 301 gaagaaccgg ctgatttcc gcaaacatg ctgaatccca aatatacatt tgatacgtc 361 gttatcggt caggaaaccg attggccaa geagegtct tggcagaag agatcacgt 421 gcgaaagett acaatcgct gttatttac gggggagtcg gactggaaa gactactta 481 atgcatgega tcggcacta tgtcatcgat cacaaccat ctgcaaaag ggttatttg 541 tcatctgaga aatttacaa tggtgcatt aactcgatc gtgacaataa agetgcgat 601 tttcgcaat gctatcgaa tgtgacgtt ctttaatag acgatatta attttagcc 661 ggaaaagaac agacgcaaga ggaattttc catacgtta atacgttc tgaagaaca 721 aagcagattg tcattccag cgaccgcet ccaaaagag tcccaacgt tgaagaaca 721 aggaagtg tttaagaa gaagcaaa gcagaagga ttgaatccc gaatgaaca 841 agaattgcg tttaagaa gaagcaaa gcagaagga ttgatccc gaatgaagt 901 atgetttata ttgccaatca gatcgacaa aacacgg ggcattaac 901 agggttgcg catattcttc cttgatcaat aaaggacatta acgccgatc ggcggatacg	
61 aageccaget ttgaaacatg gatgaaateg acaaaggeee atteattgea gggegataeg 121 etgateatea eegeacegaa egagttegee agagaetgee ttgaateaag ataeetgeae 181 etgategeeg ataegateta tgatetgaea ggagaagaat tgageattaa attegteat 241 eeteagaee aaaatgaaga agatttatg eeaaagtee caataaaata ttgetaat 301 gaagaaeegg etgatttee geaaaeatg etgaateeea aatataeatt tgateegte 361 gttateggt eaggaaaeeg attegeeea geagegtett tggeagtege tgaageeeg 421 gegaaagett acaateeget gttatttae gggggagteg gaettgaaa gaeteaetta 481 atgeatgega teggeeata tgteategaa eggeggetet tggeaaaag ggttatteg 541 teatetgaga aattaeaa tgatgaegt eacaateea etgeaaaaa agetgegat 601 ttegeaate getategeaa tgteaegt etttaatag acgatatea attttagee 661 ggaaaagae agaegeaaga ggaatttte cataegtta attegete tgaagaaea 721 aageagattg teatteeag eggeegee ecaaagaga teeeaaeget tgaageegt 841 agaatgega tttaagaa gaageaaa geagaagga tegaaaca 841 agaattgeg tttaagaa gaacgeaaga gaagataa ageetggaa 841 agaattgeg tttaagaa gaacgaaca aacaea ageetgga tgaagaea 961 aggettge eatatette ettgateag aaageagaa aegeeggae tgaageega geetgaaaea 961 aggetgee catatette ettgatea aaaagagagae ageeggaeta	
121 ctgatcatca ccgcaccgaa cgagttcgcc agaggattggc ttgaatcaag atacctgcac 181 ctgatcgccg atacgatcta tgatctgaca ggagaagaat tgagcattaa atttgtcatt 241 cctcagaatc aaaatgaaga agattttatg ccaaagtctc caatcaaaaa aatgtcgaaa 301 gaagaaccgg ctgattttcc gcaaaacatg ctgaatccca aatatacat tgatacgttc 361 gtatcggtt caggaaaccg attcgccac gcagggtct tggcagtgge tgaagcccg 421 gcgaaagctt acaatccgct gttatttac gggggggtgg gacttggaaa gactactta 481 atgcatgcga tcgggcatta tgtcatcgat cacaatccat ctgcaaaag ggttatttg 541 tcatctgaga aatttacaaa tgagttcatt aactcgatcc gtgacaataa agctgtcgat 601 tttcgcaatc gctatcgaaa tgttgacgt cttttaatag acgaatatca attttagcc 661 ggaaaagaat gcagcagaa ggaattttc catacgtta atacgctgc tgaagaaca 721 aagcagattg tcatttccag cgaccggcct ccaaaagaga tcccaccta tagaagaccgt 781 ttggcctcc gtttgaatg gggatgaa gaagagaa ctgaacacg tgaagaaca 841 agaattgcga tttaagaa gaagcaaaa gcagaagga ttgatatcc gaatgaagtc 901 atgctttata ttgccaatca gatcgacga aatatcaagg ggcattaatc 961 agggttgtcg catattcttc cttgatcaat aaagacatta acgccgatc ggctgcaa	
181 ctgatcgccg atacgatcta tgatctgaca ggagaagaat tgagcattaa atttgtcatt 241 cctcagaatc aaaatgaaga agatttatg ccaaagtctc caatcaaaaa aatgtcgaaa 301 gaagaaccgg ctgatttcc gcaaacatg ctgaatccca aatatacatt tgatacgttc 361 gttatcggtt caggaaccg attcgcccac gcagcgtctt tggcagtgge tgaagccccg 421 gcgaaagctt acaatccgct gttatttac ggggggggg gacttgggaa gactactta 481 atgcatgcga tcgggcacta tgtcatcgat cacaatccat ctgcaaaagt ggttatttg 541 tcatctgaga aatttacaaa tgagttcatt aactcgatcc gtgacaataa agctgtcgat 601 tttcgcaatc gctatcgaaa tgttgacgtt ctttaatag acgatattca attttagcc 661 ggaaaagaac agacgcaaga ggaattttc catacgtta atacgctgca tgaagaaca 721 aagcagattg tcatttccag cgaccggcct ccaaaagaga tcccacet tgaagaaca 841 agaattgcga ttttaagaa gaagcaaaa gcagaagga ttgaatccc gaatgaagtc 901 atgcttata tttcagaaa gacgcaaga gattatca actgacgg gagtagaac 961 agggttgtcg catattcttc cttgatcaat aaagcagta acggcgattaatc	
241 cctcagaatc aaaatgaaga agatttatg ccaaagtctc caatcaaaaa aatgtcgaaa 301 gaagaaccgg ctgatttcc gcaaaacatg ctgaatccca aatatacatt tgatacgttc 361 gttatcggtt caggaaaccg attcgccac gcagcgtctt tggcagtgge tgaagcccg 421 gcgaaagctt acaatccgct gttatttac ggggggggtg gacttggaaa gactactta 481 atgcatgcga tcgggcacta tgtcatcgat cacaatccat ctgcaaaagt ggttattg 541 tcatctgaga aatttacaaa tgagttcatt aatcgatcc gtgacaataa agctgcgat 601 tttcgcaatc gctatcgaaa tgttgacgtt cttttaatag acgatattca attttagcc 661 ggaaaagaac agacgcaaga ggaattttc catacgtta atacgetgca tgaagaaca 721 aagcagattg tcattccag cgaccggcet ccaaaagaga tcccaage tgaagaaca 841 agaattgcga ttttaagaa gaagcaaaa gcagaagga ttgatatcc gaatgaagtc 901 atgcttata ttgccaatc gatcgacag aatatccag tgaagaagg ggcattaatc 961 agggttgtcg catattcttc cttgatcaat aaggcgtag ggcagtaa	
301 gaagaaccgg ctgattttcc gcaaaacatg ctgaatccca aatatacatt tgatacgttc 361 gttatcggtt caggaaaccg attegeceae geagegtett tggeagtgge tgaageeeg 421 gegaaagett acaatceget gttatttae gggggagteg gaettggaa gaetcaetta 481 atgeatgega tegggeaeta tgteategat caeaatceat etgeaaaagt ggttatttg 541 teatetgaga aatttacaaa tgagtteatt aactegatee gtgaeaataa agetgegat 601 tttegeaate getategaaa tgttgaegtt etttaatag aegatattea attettagee 661 ggaaaagaac agaegeaaga ggaatttte cataegtta ataegetgea tgaagaaca 721 aageagattg teattteeag egaeeggeet eeaaaagag teeeaaget tgaagaaca 841 agaatgega tttaagaa gaageaaag geagaaggae ttgatacce gaatgaagte 901 atgettata tgeeaate gategeaeg aataccag ageagaagg ggeattaate 961 agggttgeg catattette ettgateaa aaggeatta acegeegte ggetgeaaa	
361 gttateggit eaggaaaceg attegeeea geagegtett tggeagtgge tgaageeeg 421 gegaaagett acaateeget gttatttae gggggagteg gaettggaaa gaeteaetta 481 atgeatgega tegggeaeta tgteategat caeaateeat etgeaaaagt ggttatttg 541 teatetgaga aattacaaa tgagteatt aaetegatee gtgacaataa agetgegat 601 ttegeaate getategaaa tgttgaegtt etttaatag aegatatea attettagee 661 ggaaaagaac agaegeaaga ggaatttte cataegtta ataegetgea tgaagaaca 721 aageagattg teattteeag egaeeggeet eeaaagaga teeeaaget tgaagaaca 841 agaattgega tttaagaa gaageaaag gegatgate aetgeaate egeeteega ttgaagaaca 841 agaattgega ttgeeaata gategaega gaagtage ggeattaate 901 atgetttata ttgeeatee gategaega gaetgaega ageeggagg geetgaa 961 agggttgteg catattette ettgateaat aaageaetta acgeegaet ggetgetga	
421 gegaaagett acaateeget gtttatttae gggggagteg gaettggaaa gaeteaetta 481 atgeatgega tegggeaeta tgteategat caeaateeat etgeaaaagt ggtttatttg 541 teatetgaga aatttaeaaa tgagtteatt aaetegatee gtgaeaataa agetgtegat 601 tttegeaate getategaaa tgttgaegtt ettttaatag aegatattea attttagee 661 ggaaaagaac agaegeaaga ggaatttte cataegtta ataeegetgea tgaagaaaca 721 aageagattg teatteeag eggaeggeet eeaaagaga teesaaeget tgaagaeeag 781 ttggeetee gtttgaagt gggattgate aetgaeatea egeeteetgaaaca 841 agaattgega ttttaagaaa gaageaaaa geegaaggag ttgataece gaatgaagte 901 atgetttata ttgeeaatea gategaeage aatateaggg agetggaagg ggeattaate 961 agggttgteg eatattette ettgateaat aaagaeatta acgeegtet ggetgetgaa	
481 atgcatgcga tegggeacta tgteategat caeaateeat etgeaaaagt ggtttatttg 541 teatetgaga aatttacaaa tgagtteatt aaetegatee gtgacaataa agetgtegat 601 tttegeaate getategaaa tgttgaegtt ettttaatag aegatattea atttttagee 661 ggaaaagaae agaegeaaga ggaatttte cataegttta ataegetgea tgaagaaea 721 aageagattg teatteeag egaeeggeet eeaaagagga teeeaaegt tgaagaeeg 781 ttgegetee gttttgaatg gggatgate aetgaeatea egeeteetgaaaea 841 agaattgega ttttaagaaa gaaageaaag gaagagga etgatatee gaatgaagte 901 atgetttata ttgeeaatea gategaeage aatateaggg agetggaagg ggeattaate 961 agggttgteg eatattette ettgateaat aaagaeatta aegeegate ggetgetgaa	
541 teatetgaga aatttacaaa tgagtteatt aactegatee gtgacaataa agetgtegat 601 tttegeaate getategaaa tgttgaegtt ettttaatag aegatattea atttttagee 661 ggaaaagaae agaegeaaga ggaatttte cataegttta ataegetgea tgaagaaeaa 721 aageagattg teattteeag egaeeggeet eeaaaagaga teeeaaeget tgaagaeeg 781 ttgegeteee gttttgaatg gggattgate aetgaeatea egeeteetgaagae 841 agaattgega ttttaagaaa gaaageaaaa geagaaggae ttgataeee gaatgaagte 901 atgetttata ttgeeaatea gategaeage aataeeagg ggeattaate 961 agggttgteg eatattette ettgateaat aaagaeatta aegeegate ggetgetgaa	
601 tttegeaate getategaaa tgttgaegtt ettttaatag aegatattea atttttagee 661 ggaaaagaac agaegeaaga ggaatttte cataegtta ataegetgea tgaagaaca 721 aageagattg teattteeag egaeeggeet eeaaaagaga teeeaaeget tgaagaeeg 781 ttgegeteee gttttgaatg gggattgate aetgaeatea egeeteega tetggaaaea 841 agaattgega ttttaagaaa gaaageaaa geagaaggae ttgatatee gaatgaagte 901 atgetttata ttgeeaatea gategaeage aataeeagg agetggaagg geattaate 961 agggttgeg catattette ettgateaat aaagaeatta aegeegate ggetgetgaa	
661 ggaaaagaac agacgcaaga ggaatttttc catacgttta atacgctgca tgaagaaaca 721 aagcagattg tcatttccag cgaccggcct ccaaaagaga tcccaacgct tgaagaccgt 781 ttgcgctccc gttttgaatg gggattgatc actgacatca cgcctcctga tctggaaaca 841 agaattgcga ttttaagaaa gaaagcaaaa gcagaaggac ttgatatccc gaatgaagtc 901 atgctttata ttgccaatca gatcgacagc aatatcaggg agctggaagg ggcattaatc 961 agggttgtcg catattcttc cttgatcaat aaagacatta acgccgatct ggctgctgaa	
721 aagcagattg teattteeag egaeeggeet eeaaagaga teecaaeget tgaagaeegt 781 ttgegeteee gtttgaatg gggattgate aetgaeatea egeeteetga tetggaaaea 841 agaattgega ttttaagaaa gaaagcaaaa geagaaggae ttgatateee gaatgaagte 901 atgetttata ttgeeaatea gategaeage aatateaggg agetggaagg ggeattaate 961 agggttgteg eatattette ettgateaat aaagaeatta aegeegatet ggetgetgaa	
781 ttgcgetece gttttgaatg gggattgate actgacatea egeeteetga tetggaaaca 841 agaattgega ttttaagaaa gaaageaaaa geagaaggae ttgatateee gaatgaagte 901 atgetttata ttgeeaatea gategaeage aatateaggg agetggaagg ggeattaate 961 agggttgteg catattette ettgateaat aaagaeatta aegeegatet ggetgetgaa	
841 agaattgega tittaagaaa gaaageaaaa geagaaggae tigatateee gaatgaagte 901 atgetitata tigeeaatea gategaeage aatateaggg ageiggaagg ggeattaate 961 agggitgieg catatteite eitgateaat aaagaeatta aegeegatei ggeigeigaa	
901 atgetttata ttgeceatea gategaeage aatateaggg agetggaagg ggeattaate 961 agggttgteg catattette ettgateaat aaagaeatta aegeegatet ggetgetgaa	
1021 getttgaaag atateattee ttetteaaag eegaaagtea ttaegateaa agaeateeaa	
1081 agaatcgtcg gccagcagtt taatatcaag ctggaggatt tcaaggcgaa gaaacggaca	
1141 aaatcggtgg cttttccgcg gcagatcgct atgtatctat caagagaaat gacagattct	
1201 tetetteega agateggega agaatttgge ggaegegaee acaegaeggt cateeatgee	
1261 catgagaaaa tatcaaaact gctgagcgat gatgaacagc ttcagcagca gattaaagaa	
1321 attaaagagc agctgagata a	
	>

http://www.ncbi.nlm.nih.gov/





### Sequencing a Genome

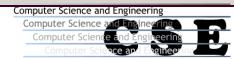
Genomes are determined using a technique known as *shotgun sequencing*.

Computer scientists have played an important role in developing algorithms for assembling such data.

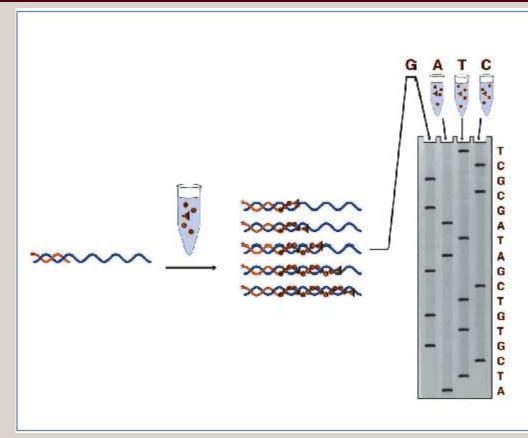
It's kind of like putting together a jigsaw puzzle with millions of pieces (a lot of which are "blue sky"). Whole Genome Shotgun Sequencing Method Genomic DNA Genomic DNA Sequence Each Fragment with Shotgun Approach CONTROCTOBORCANCERE C

 $http://occawlonline.pearsoned.com/bookbind/pubbooks/bc\_mcampbell\_genomics\_1/medialib/method/shotgun.html$ 





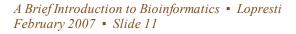
## Reading DNA



#### This is known as Sanger sequencing.

http://www.apelex.fr/anglais/applications/sommaire2/sanger.htm http://www.iupui.edu/~wellsctr/MMIA/htm/animations.htm Gel electrophoresis is process of separating a mixture of molecules in a gel media by application of an electric field. In general, DNA molecules with similar lengths will migrate same distance.

First "starve" DNA copy reaction at each base: *A*, *C*, *G*, *T*. Then run gel and read off sequence: *ATCGTG* ...

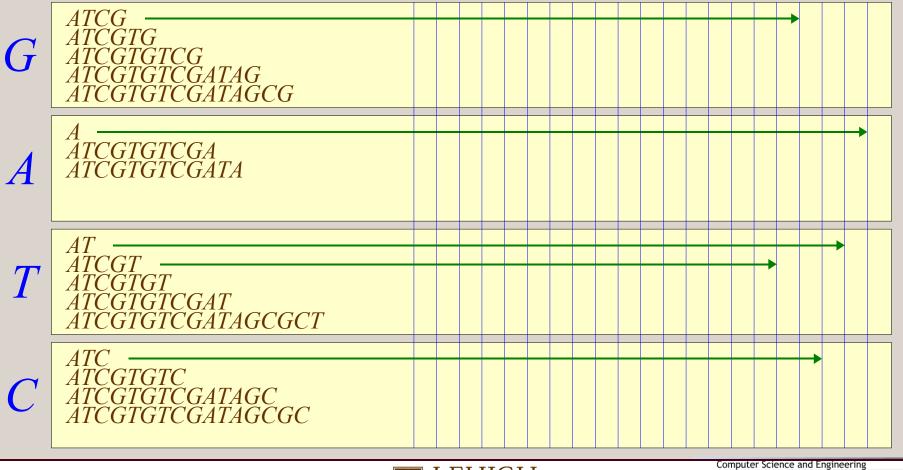




Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering

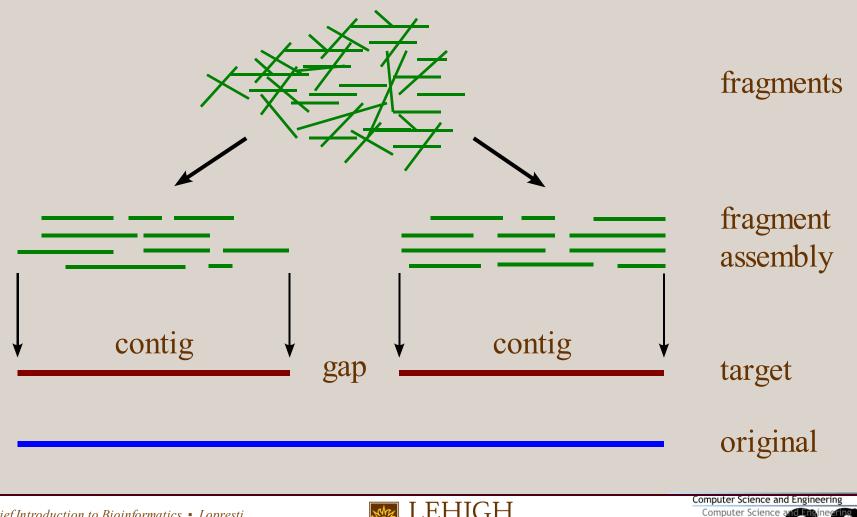
### Reading DNA

#### Original sequence: ATCGTGTCGATAGCGCT





Computer Science and Eng Computer Science and E

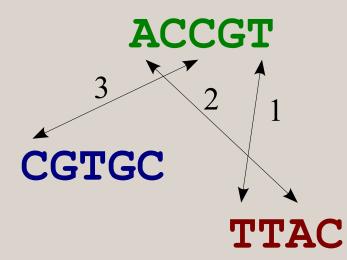




Computer Scien

A simple model of DNA assembly is the *Shortest Supersequence Problem*: given a set of sequences, find the shortest sequence *S* such that each of original sequences appears as subsequence of *S*.

Look for overlap between prefix of one sequence and suffix of another:

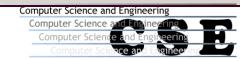


--ACCGT--

- ---CGTGC
- TTAC----

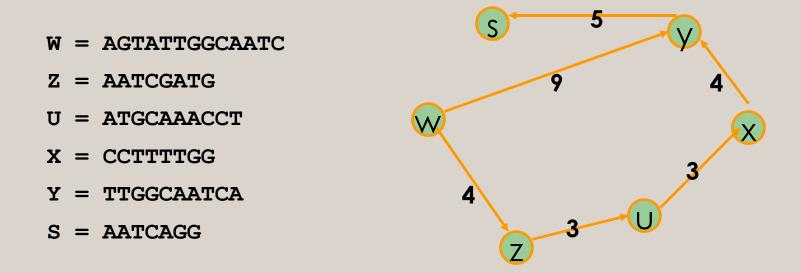
TTACCGTGC





#### Sketch of algorithm:

- Create an overlap graph in which every node represents a fragment and edges indicate overlap.
- Determine which overlaps will be used in the final assembly: find an optimal spanning forest in overlap graph.

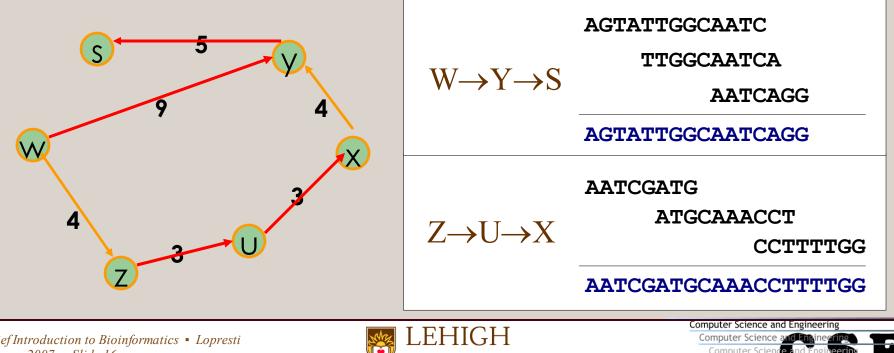




Computer Science and Engineering Computer Science and Engineering

Computer Sc

- Look for paths of maximum weight: use greedy algorithm to select edge with highest weight at every step.
- Selected edge must connect nodes with in- and out-degrees  $\leq 1$ .
- May end up with set of paths: each corresponds to a contig.





What's the problem? Google for biologists ...

- Given new DNA or protein sequence, biologist will want to search databases of known sequences to look for anything similar.
- Sequence similarity can provide clues about function and evolutionary relationships.
- Databases such as GenBank are far too large to search manually. To search them efficiently, we need an algorithm.

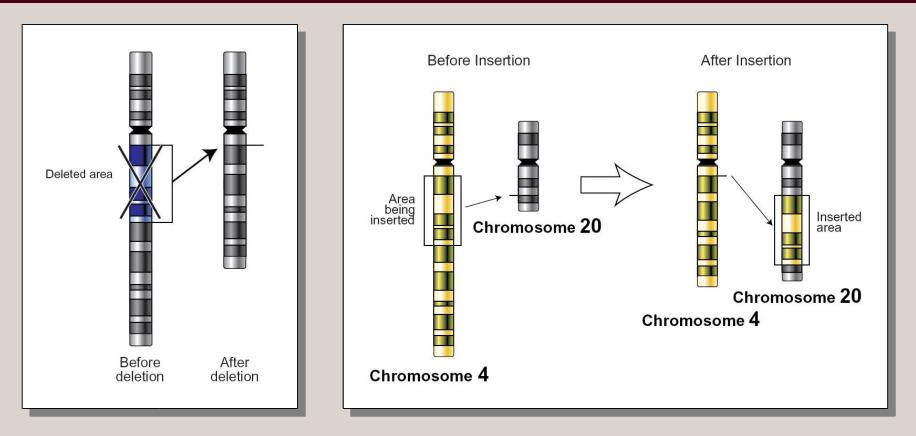
Shouldn't expect exact matches (so it's not really like google):

- Genomes aren't static: mutations, insertions, deletions.
- Human (and machine) error in reading sequencing gels.





### Genomes Aren't Static



#### Sequence comparison must account for such effects.

http://www.accessexcellence.org/AB/GG/nhgri\_PDFs/deletion.pdf

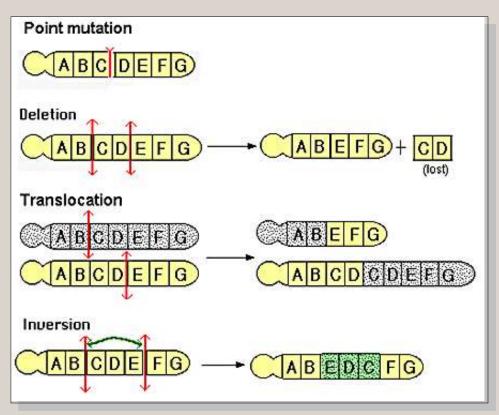
http://www.accessexcellence.org/AB/GG/nhgri\_PDFs/insertion.pdf





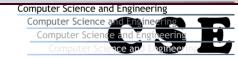
### Genomes Aren't Static

#### Different kinds of mutations can arise during DNA replication:



http://www.accessexcellence.org/AB/GG/mutation.htm



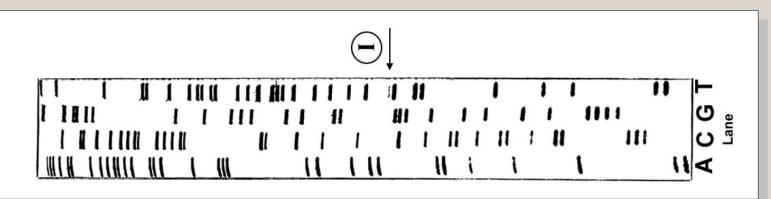


### The Human Factor

In addition, errors can arise during the sequencing process:

"...the error rate is generally less than 1% over the first 650 bases and then rises significantly over the remaining sequence."

A hard-to-read gel (arrow marks location where bands of similar intensity appear in two different lanes):

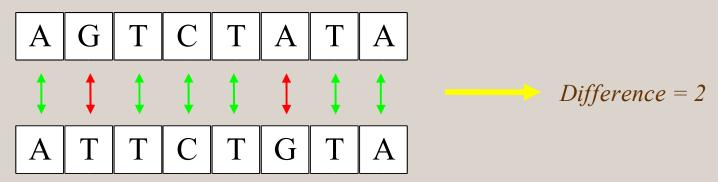


 $http://hshgp.genome.washington.edu/teacher\_resources/99-studentDNASequencingModule.pdf$ 

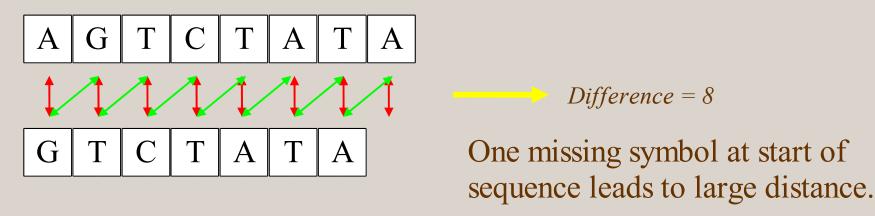




Why not just line up sequences and count matches?



Doesn't work well in case of deletions or insertions:

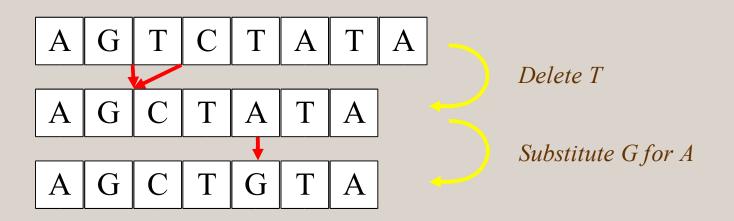




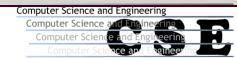
Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering

Instead, we'll use a basic technique known as dynamic programming.

- Model allows three basic operations: delete a single symbol, insert a single symbol, substitute one symbol for another.
- Goal: given two sequences, find the shortest series of operations needed to transform one into the other.

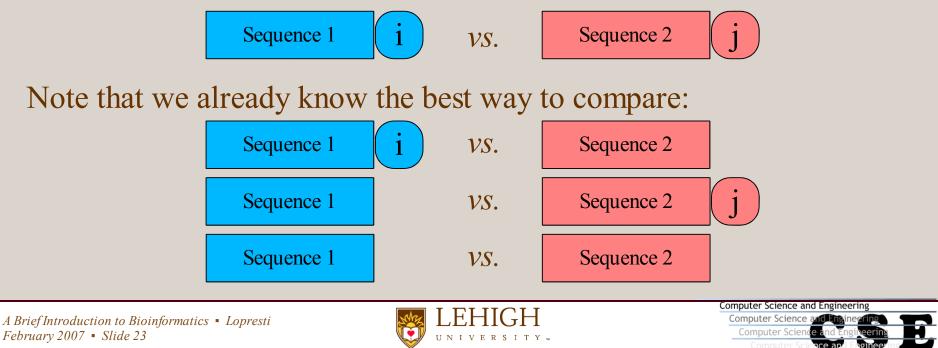






How can we determine optimal series of operations?

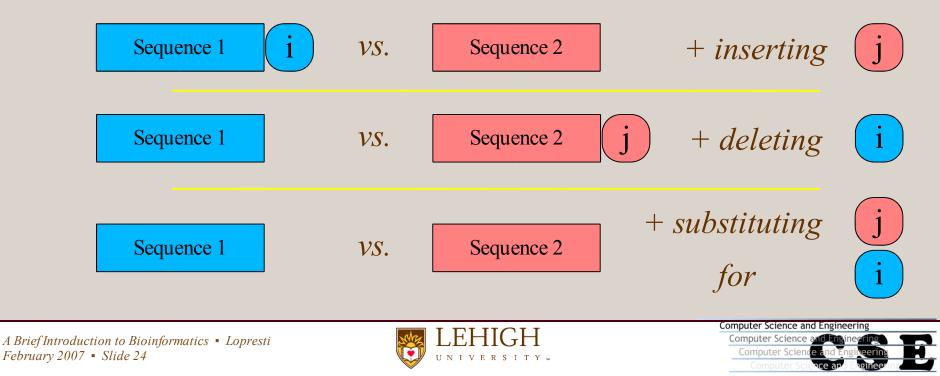
- Approach is to build up longer solutions from previously computed shorter solutions.
- Say we want to compute solution at index *i* in first sequence and index *j* in second sequence:



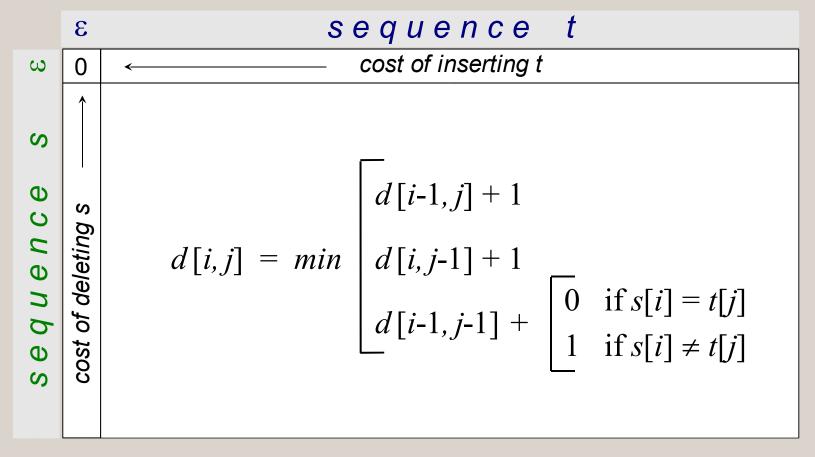
#### So, best way to do this comparison:

Sequence 1 i VS. Sequence 2 j

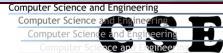
#### Is best choice from following three cases:



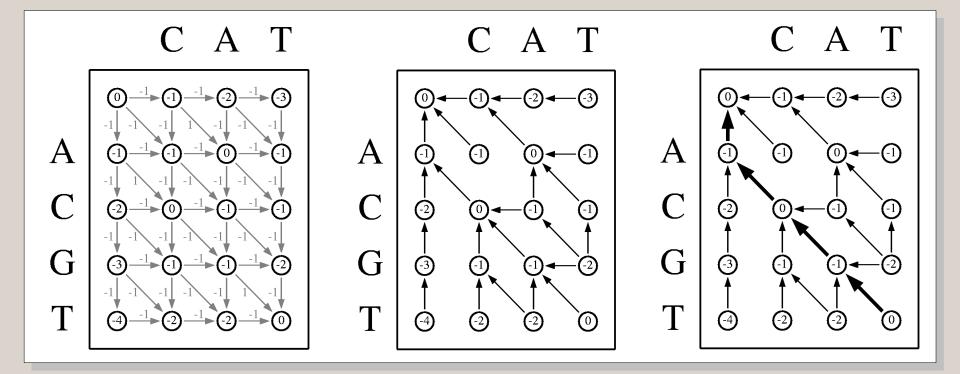
#### Normally, this computation builds a table of distance values:



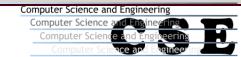




#### By keeping track of optimal decision, we can determine operations:

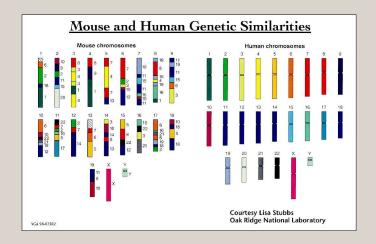






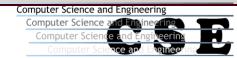
### Genome Rearrangements

Recall what we saw earlier:



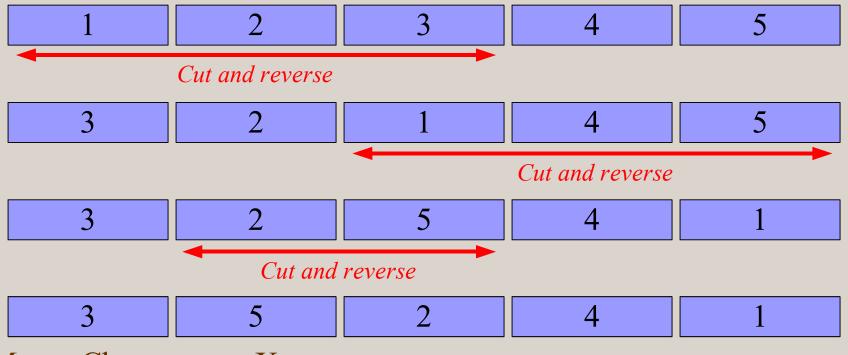
- 99% of mouse genes have homologues in human genome.
- 96% of mouse genes are in same relative location to one another.
- Mouse genome can be broken up into 300 *synteny blocks* which, when rearranged, yield human genome.
- Provides a way to think about evolutionary relationships.





### **Reversal Distance**

#### Human Chromosome X



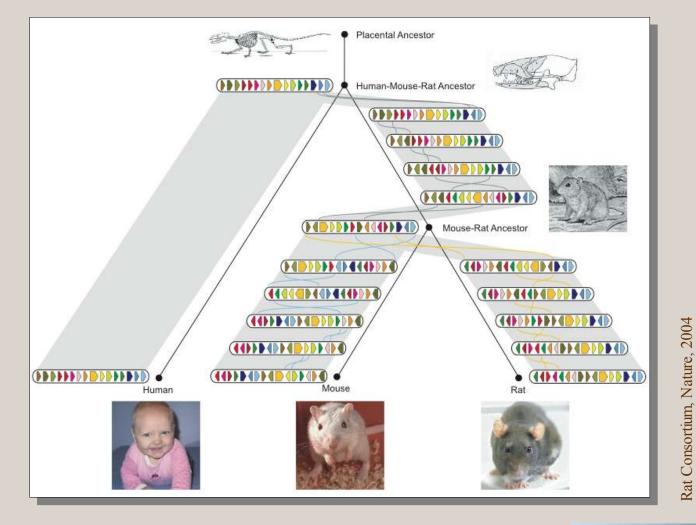
Mouse Chromosome X

Reversal distance is the minimum number of such steps needed.





### History of Chromosome X



A Brief Introduction to Bioinformatics • Lopresti February 2007 • Slide 29

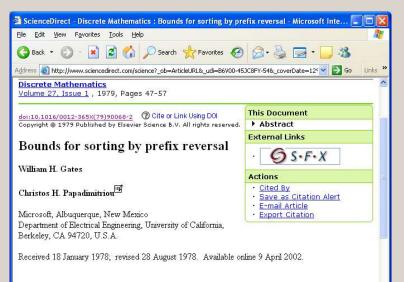


Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering

### Interesting Sidenote

Early work on a related problem, sorting by prefix reversals, was performed in 1970's by Christos Papadimitriou, a famous computer scientist now at UC Berkeley, and one "William H. Gates" ...





#### Abstract

For a permutation  $\sigma$  of the integers from 1 to n, let  $f(\sigma)$  be the smallest number of prefix reversals that will transform  $\sigma$  to the identity permutation, and let f(n) be the largest such  $f(\sigma)$  for all  $\sigma$  in (the symmetric group)  $S_n$ . We show that  $f(n) \leq (5n+5)/3$ , and that  $f(n) \geq 17n/16$  for n a multiple of 16. If, furthermore, each integer is required to participate in an even number of reversed prefixes, the corresponding function g(n) is shown to obey  $3n/2-1\leq g(n)\leq 2n+3$ .

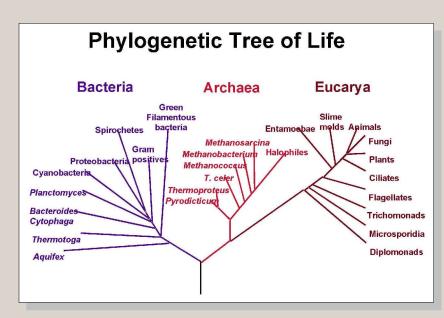
#### Yes, that Bill Gates ...



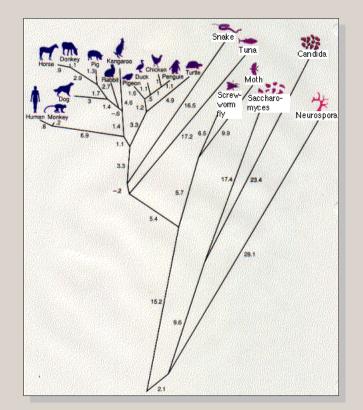


### Building the "Tree of Life"

Scientists build phylogenetic trees in an attempt to understand evolutionary relationships. Reversal distance is often used here.



Note: these trees are "best guesses" and certainly contain some errors!



Computer Science and Engineering

Computer Science

Computer Scie

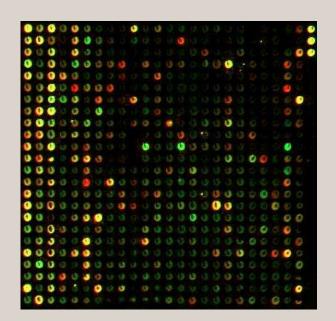
http://en.wikipedia.org/wiki/Phy logenetic\_tree http://users.rcn.com/jkimball.ma.ultranet/Biology Pages/T/T axonomy .html



### DNA Microarrays

- Allows simultaneous measurement of the level of transcription for every gene in a genome (gene expression).
- Differential expression, changes over time.
- Single microarray can test ~10k genes.
- Data obtained faster than can be processed.
- Want to find genes that behave similarly.
- A pattern discovery problem.

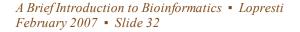
green = repressed red = induced



Computer Science and Engineering

Computer Science

Computer Sc





### **DNA** Microarrays

*K-means clustering* is one way to organize this data:

- Given set of *n* data points and an integer *k*.
- We want to find set of *k* points that minimizes the mean-squared distance from each data point to its nearest cluster center.

### Sketch of algorithm:

- Choose *k* initial center points randomly and cluster data.
- Calculate new centers for each cluster using points in cluster.
- Re-cluster all data using new center points.
- Repeat second two steps until no data points are moved from one cluster to another or some other convergence criterion is met.





### Clustering Microarray Data

0

0

0

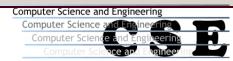
0

- Pick k = 2 centers at random.
- Cluster data around these center points.

• Re-calculate centers based on current clusters.

From "Data Analysis Tools for DNA Microarrays" by Sorin Draghici.





### Clustering Microarray Data

• Re-cluster data around new center points.

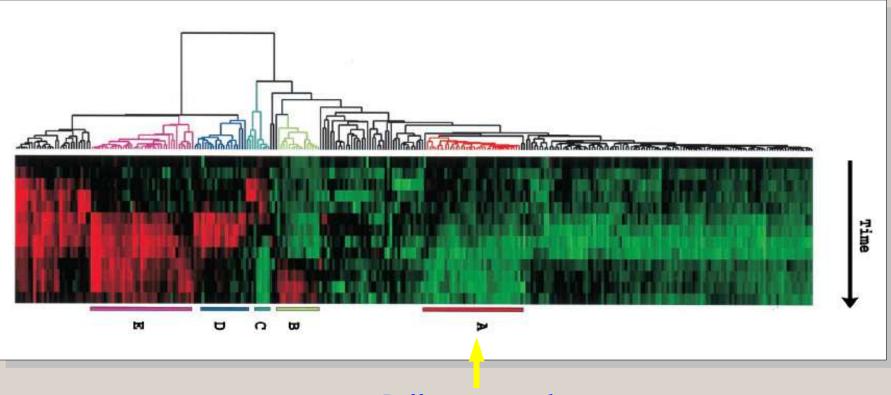
• Repeat last two steps until no more data points are moved into a different cluster.

From "Data Analysis Tools for DNA Microarrays" by Sorin Draghici.





### Example of Hierarchical Clustering



#### Different genes that express similarly

From "Cluster analysis and display of genome-wide expression patterns" by Eisen, Spellman, Brown, and Botstein, Proc. Natl. Acad. Sci. USA, Vol. 95, pp. 14863–14868, December 1998

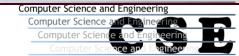




### Why Study Bioinformatics?

- Still many urgent open problems ⇒ lots of opportunities to make fundamental contributions (and become rich and famous).
- Stretch your creativity and problem-solving skills to the limit.
- Join a cross-disciplinary team work with interesting people.
- Participate in unlocking the mysteries of life itself.
- Make the world a better place.





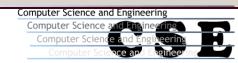
### CSE Course in Bioinformatics

In CSE 308/408, we study algorithms for:

- Sequence comparison & alignment (pairwise & multiple).
- Sequence assembly (shotgun sequencing).
- Physical mapping of DNA.
- Constructing phylogenetic (evolutionary) trees.
- Computing genome rearrangements.
- DNA microarray analysis.
- DNA computing.

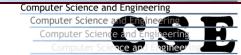
### Materials @ http://www.cse.lehigh.edu/~lopresti/courses.html Questions: dal9@lehigh.edu





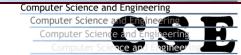
# Thank you!





# Backup Slides





### **Bioinformatics and Computer Science**

Recall that bioinformatics is the application of techniques from computer science to problems from biology.

Particularly relevant subfields:

- pattern recognition (classifying unknown inputs),
- image processing (machine vision),
- databases (efficient storage and retrieval),
- data mining (extracting knowledge from vast datasets),
- graphics & visualization (assists in data analysis),
- robotics (automating DNA microarray experiments).





#### Simple example:

		q	u	i	c	k
	0	1	2	3	4	5
h	1	1	2	3	4	5
a	2	2	2	3	4	5
c	3	3	3	3	3	4
k	4	4	4	4	4	3

So the distance between "hack" and "quick" is 3.

*This corresponds to one insertion and two substitutions.* 

If one sequence has length M and other has length N, then table has  $M^*N$  entries. We say the time complexity is O(MN).



