Class: Biol/CS123A **Name:** Zayd Hammoudeh

Assignment: #4

Problem: #1 / |

Problem: #2 / |

Problem: #3 / |

Problem: #4 / |

Problem: #5 / |

Problem: #6 / |

Problem: #7 / |

Grade: Total / |

Problem #1

Note: The page numbers below are for my studying later on and can be ignored during grading.

Chapter #4 (Continued):

1.1) Upon scoring pairs of aligned amino acids, why are pairs such as isoleucine and leucine scored more highly than other pairs such as isoleucine and lysine?

Not all amino acid pairs are as equally likely to occur in homologous proteins. For example, isoleucine and leucine have similar properties. In contrast, isoleucine and lysine have quite different properties and are rarely found in corresponding positions in homologous protein sequences. These varying degrees of similarity and dissimilarity necessitate a more complex metric of assessing similarity be used. As such, each aligned pair of amino acids is given a numerical score based on the probability of the relevant change occurring during evolution. (See page 80)

1.2) What is the minimum percentage identity that can reasonably be accepted as significant when comparing two protein sequences? Explain making sure to include what is meant by the twilight zone in your answer.

Burkhard Rost found that 90% of amino acid sequence pairs with identity scores greater than 30% were structurally similar proteins. Due to that, 30% is generally taken as the threshold for an initial presumption of homology. Below 25%, only 10% of aligned amino acid sequence pairs were structurally similar. The region between 20% and 30% identity in aligned amino acid sequences is known as the **twilight zone**. When a sequence pair falls in that region, homology may exist but cannot be reliably assumed in the absence of other evidence. The **midnight zone** is when two aligned sequences have less than 20% identity. (See page 81).

1.3) When we are deciding on which substitution matrix to use, is there such a thing as one correct scoring scheme for all circumstances? Explain.

There is not one scoring system that is ideal for all circumstances. For example, there is a wide range of variation in the similarity of sequences, from almost complete identity to only a few percent of similarity. In some cases, the goal is to align and score closely related sequences while in other cases the goal is identify very distant evolutionary relationships reliably. In either case, a different substitution matrix is needed. (See page 82)

1.4) The rows in the substitution matrices of Figure 4.4 on page 83 are color coded. Explain what each color represent and write one or two sentences on each property.

There are six colors in the substitution scoring matrix. The color shading indicates the different physiochemical properties of the residues. Yellow indicates small, polar amino acids, specifically Cysteine (C), Serine (S), and Threonine (T). Polar amino acids are charged and tend to be hydrophilic.

Small and nonpolar amino acids are in white and include Proline (P), Alanine (A), and Glycine (G). Their functional groups consist primarily of hydrocarbons and are uncharged. They are incapable of hydrogen bonding.

The amino acids in red are polar or acidic; this category includes Asparagine (N), Aspartic acid (D), Glutamic acid (E), and Glutamine (Q). Aspartic acid and glutamic acids have acidic side chains at neutral pH. Their side chains have carboxyylic acid groups. Asparagine and glutamine have amino groups as part of their side chain in addition to their hydrocarbons.

Basic amino acids are shown in blue and include Histidine (H), Arginine (R), and Lysine (K). Their side chain has an NH_x (Ammonia) molecule and is generally positively charged.

Green amino acids are large and hydrophobic (not attracted to water) and includes Methionine (M), Isoleucine (I), Leucine (L) and Valine (V). Their side changes are generally composed of CH_x molecules. Hydrophobic amino acids are generally uncharged.

Orange represents the aromatic amino acids and includes Phenylalanine (F), Tyrosine (Y), and Tryptophan (W). Aromatic amino acids have an aromatic (unusually stable, flat) ring of atoms.

1.5) Explain each of the following terms.

a. **PAM** – Point accepted mutation. It refers to the amount of point (i.e. single nucleotide) mutations that are retained in a sequence per 100 residues. Scoring matrices that rely on point accepted mutations are known as PAM matrices; they are also referred to as Dayhoff mutation data matrices. It was derived from a comparatively small number of protein families in the 1970s.

b. **BLOSUM** – An acronym for Block Substitution. It was developed in the early 1990's using local sequence alignments. It utilized a large set of highly conserved short amino acid sequences (called Blocks) and determined substitution frequencies for all possible pairs of amino acids.

c. **PET91** – A new generation of Dayhoff-mutation matrices. It was updated by Jones et. Al to include more protein families.

d. **STR** – Amino acid substitution matrix that includes information from known protein structures. Since protein structure is more conserved than protein sequence, most distantly related proteins can be compared using this approach.

e. **SLIM** – Score Matrix Leading to Intra-Membrane. Designed specifically for membrane proteins where the amino acid composition and the selective forces for acceptable mutations are different than for soluble proteins.

f. **PHAT** – Predicted Hyrdophobic and Transmembrane Matrix. Designed specifically for membrane proteins where the amino acid composition and the selective forces for acceptable mutations are different than for soluble proteins. (See page 84)

1.6) What are the two findings that were obtained from structural analysis of the sequence of amino acids with respect to insertions and deletions of amino acids?

Structural analysis has shown that few insertions and deletions occur in sequences of structural importance. What is more, insertions tend to be several residues long rather than a single residue long. (See page 85)

1.7) When should we set a high gap penalty, and when should we set a low gap penalty?

To place limits on the insertions of gaps in alignment, a **gap penalty** is subtracted from the alignment score any time a gap is added. When a new gap is inserted into a sequence, the score is penalized relatively more than when an existing gap is lengthened. The penalty to extend an existing gap is known as the **gap extension penalty**.

The insertion of a gap is intended to improve the quality of a sequence alignment. If a gap penalty is set high, then fewer gaps will be inserted into the alignment as their inclusion will drastically reduce the maximum match value. Thus, if one is searching for sequences that are a strict match for your query sequence, the gap penalty should be set high. In contrast, if one is searching for similarity between distantly related sequences, the gap penalty should be set low. (See page 85

1.8) Why should amino acids such as tryptophan have higher gap penalties (when aligned with gaps) than other amino acids such as glycine?

Some residues are more likely to be conserved since their side chains tend to be more important in determining the structure or function of a protein. An example of this is tryptophan so it receives a larger gap penalty than for example glycine. (See page 85).

1.9) What are the Smith-Waterman and Needleman-Wunsch algorithms? What are they used for? Which one is a special case of the other?

Smith-Waterman and **Needleman-Wunsch** are dynamic programming algorithms used for aligning sequences. Needleman-Wunsch is used for global sequence alignment while Smith Waterman is used in local sequence alignment. Smith Waterman is a modification of the Needleman-Wunsch algorithm. (See pages 87-88)

1.10) What do sequence conservations in multiple alignments identify?

Multiple sequence alignment is especially useful for illustrating sequence conservation throughout the aligned sequences. Such conservations over many sequences can identify amino acids that are important for function or for the structural integrity of the protein fold.

Multiple sequence alignment also provides insight into protein similarity and evolutionary relationships. (See page 90)

1.11) What do the authors mean when they claim Needleman-Wunch and Smith-Waterman are more rigorous methods? What programming technique are these methods based on?

Needleman-Wunsch and Smith-Waterman are more rigorous than many other algorithms since they are guaranteed to find the best scoring alignments between two sequences (as they are dynamic programming algorithms). Two suites of programs that use dynamic programming are FASTA and BLAST. However, sequence similarity is only run on those query sequences with sufficient similarity as found by a heuristic that is not rigorous. (See page 95)

1.12) Is BLAST a rigorous method? Is SSEARCH a rigorous method?

BLAST is not a rigorous method since it does not run dynamic programming on all sequences in the database. Instead, it relies on finding core similarity, which is defined by a window of preset size. Only those sequences that have a sufficient initial score are then fully aligned using dynamic programming. (See page 95) SSEARCH is search program based on the Smith-Waterman algorithm and is rigorous. It is slower than BLAST and FASTA. **SSEARCH** performs a rigorous search for similarity between a query sequence and the database. (See page 97)

1.13) Is the default gapped setting of BLAST adequate for most applications?

The gapped setting of BLAST is usually the default. It reports the best local alignment and is suitable for most applications. (See page 95).

1.14) What are the differences between blastn, blastp, and blastx?

Blastp compares an amino acid query sequence against a protein-sequence database. **Blastn** compares a nucleotide query sequence against a nucleic acid sequence database. **Blastx** compares a nucleotide query sequence translated into all six reading frames against a protein sequence database.

1.15) Consider figure 4.12 (A). Why does the caption claim that the hits above the arrow are significant while the ones below are not?

An expectation value or **E-value** is a statistical measure for estimating the significance of alignments when one alignment has been submitted for a homology search against a sequence database. The smaller the E-Value, the better the chance that the sequence that matched the query sequence is truly related. (See definition on page 738) The hits below the arrow have a much higher E-Value than those above the arrow. Quite closely related sequences have E-Values less than 10^{-20}. Given the magnitude of the E-Value, it is unlikely the query sequence is related to the sequence below the arrow.

1.16) Name two actions one could undertake to reduce the large number of hits one gets upon blasting a sequence against a database of sequences.

In many search packages, the default E-value is set to a threshold of either 0.01 or 0.001. By lowering the maximum E-value, matches above that threshold will not be included in the results. (See page 98) A query could be refined by only searching a subset of the data. For example, one could only search the newest sequences in the database or a specific genome database. (See page 100)

1.17) What are low complexity regions? Are they desirable? Why?

Low complexity regions are sequence segments that have a relatively simple structure, which is often composed of only a few different types of bases or amino acids. They are often removed from protein sequences before a database search as they can result

in misleading hits since they can artificially inflate the score and obscure biologically significant hits. An example of a low complexity would be a string of prolines or acidic amino acids. (See pages 100-101 and page 741)

1.18) Why does it make sense to resubmit a query sequence sometime after it was found to have no match in the database?

If no match is found for a query sequence, it does not necessarily mean that no there are no homologs in the database. Instead, it may mean that the similarity is too weak to be picked up by existing techniques. Techniques are continually being improved, and the amount of sequence data continues to increase. As such, it is recommended to periodically submit queries.

1.19) What is the simplest method of constructing a pattern or motif? How does the method work?

A **motif** is any conserved element of a sequence alignment (either a short sequence of contiguous residues or a more distributed pattern). The simplest way of constructing a pattern or motif is the consensus method, in which the most similar regions in a multiple sequence alignment are used to construct a pattern. Those positions in the alignment that are occupied by the same residue (or a limited subset of residues) are used to define the pattern at these positions, by specifying just the allowed residue. More sophisticated patterns can be generated using scoring tables to assess the similarity of matched amino acids. In this case, instead of just defining the pattern as requiring, for example glutamic or aspartic acid at a given position, different residues at these positions have associated scores. (See page 105)

1.20) What are logos? How are they constructed? What do the size of the letters indicate?

Logos are computed using a position-specific scoring matrix; they are a visual representation of a set of aligned sequences that indicates the position preferences as given by information theory. For each position in the sequence, a letter or set of letters is shown. The size of the letters indicates the level of conservation for a particular amino acid. The largest letters, which occupy the whole position, represent identities in the multiple sequence alignment. The colors indicate the physicochemical properties of the residues. They can be created by submitting multiple sequences to the BLOCKS program. (See page 106)

Problem #2

Go to NCBI and retrieve (data entry) with accession number Z48051.

a. **What is the name of the gene, and what is its locus?**

It is the *homo sapiens* gene for myelin oligodendrocyte glycoprotein (MOG). Its locus is 6p22.1.

b. **How many exons and introns does the gene have?**

It has 7 introns and 8 exons.

c. **The second "misc_feature" is "polymorphic (TAAA)n". What is the meaning of this entry?**

It is a tetranucleotide sequence (thymine (T) followed by three adenine (AAA)) that is repeated in the DNA sequence. It occurs from bases 16250 to 16289 (repeats 10 times). This sequence may be a genetic marker and is included for additional study.

d. **Exon #7 is from base pair 14658 to base pair 14678. What are the amino acids produced by exon 7?**

Below are the lengths of all of the exons in the coding sequence. While the protein as a whole begins in the first codon, since not all exon lengths are evenly divisible by 3, it is necessary to determine the right codon location based off the length of previous or subsequent exons.

Exon #1 Length: 88 Nucleotides (1166-1253)
Exon #2 Length: 348 Nucleotides (3274-3621)
Exon #3 Length: 114 Nucleotides (10106-10219)
Exon #4 Length: 21 Nucleotides (11597-11617)
Exon #5 Length: 21 Nucleotides (11860-11880)
Exon #6 Length: 117 Nucleotides (14238-14354)
Exon #7 Length: 21 Nucleotides (14658-14678)
Exon #8 Length: 14 Nucleotides (15129-15142)

From the length of exon #8, you are able to determine, that the last nucleotide of this exon (#7) is used to complete the first codon in exon #8. Hence, the nucleotide and amino acid sequences are below. Note for the first codon, the first base is the last base from codon #6 (G – shown in red). What is more, the last two bases in the sequence are the first two bases in exon #8 (GA – shown in red).

GGG	CAA	UUC	CUU	GAA	GAG	CUA	C**GA**
Gly (G)	Gln (Q)	Phe (F)	Leu (L)	Glu (E)	Glu (E)	Leu (L)	Arg (R)

e. **Exon #5 is from base pair 11860 to base pair 11880. What are the amino acids produced by exon #5?**

Similar to part D, the last nucleotide of this exon (#5) is used to complete the first codon in exon #6. Hence, the nucleotide and amino acid sequences are below. Note for the first codon, the first base is the last base from codon #4 (G – shown in red). What is more, the last two bases in the sequence are the first two bases in exon #8 (AU – shown in red).

GAG	AAU	CUC	CAC	CGG	ACU	UUU	G**AU**
Glu (E)	Asn (N)	Leu (L)	His (H)	Arg (R)	Thr (T)	Phe (F)	Asp (D)

Problem #3

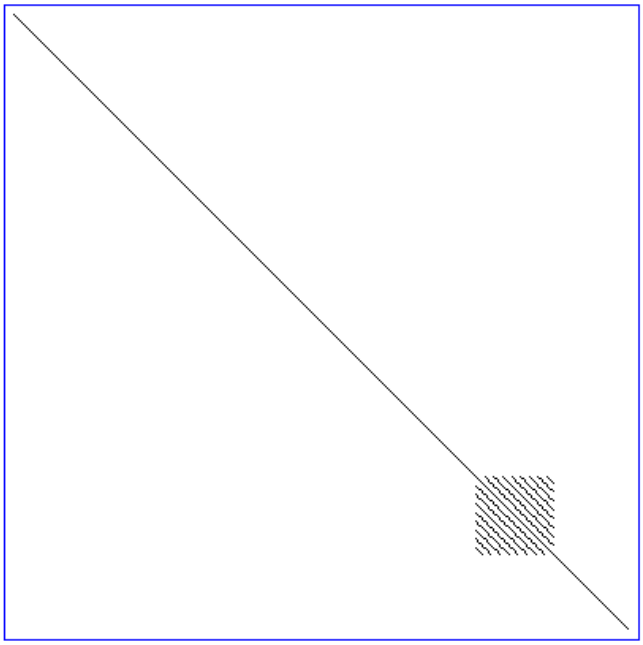

Figure 1 – Dot Plot of the Last 285 Nucleotides of Exon 8 of the MOG Protein with Mismatch Limit = 0

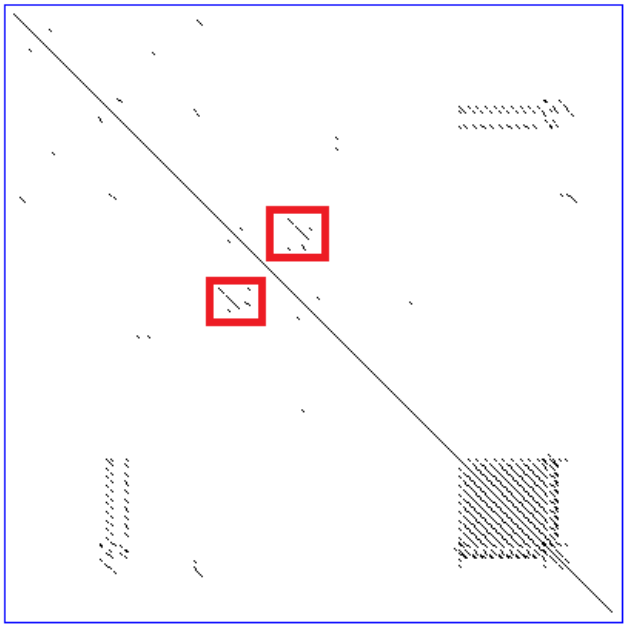

Figure 2 – Dot Plot of the Last 285 Nucleotides of Exon 8 of the MOG Protein with Mismatch Limit = 2

In both dot plots (i.e. mismatch limit equal to 0 or to 2), the tetranucleotide sequence (TAAA)n (from question #2) is evident. It is the square region of matches in the bottom right corner of the dot plot. There are also some repeat sequences that are evident when the mismatch limit is increased to two. Two primary examples are circled in red.

Problem #4

The beta globin gene ends with the following nucleotide sequence:

… CTG GCC CAC AAG TAT CAC TAA

a. Translate the above sequence and give the resulting amino acids.

…	CUG	GCC	CAC	AAG	UAU	CAC	UAA
…	Leu (L)	Ala (A)	His (H)	Lys (K)	Tyr (Y)	His (H)	STOP

b. Write a nucleotide sequence of a single base change providing a silent mutation in the region.

…	CUU	GCC	CAC	AAG	UAU	CAC	UAA
…	Leu (L)	Ala (A)	His (H)	Lys (K)	Tyr (Y)	His (H)	STOP

A **silent mutation** is any mutation that does not affect the resulting amino acid sequence. In the above sequence, the first codon was changed from CUG to CUU, both of which map to Leucine.

c. Write a nucleotide sequence and translation to an amino acid sequence of a single base change producing a missense mutation.

…	GUG	GCC	CAC	AAG	UAU	CAC	UAA
…	Val (V)	Ala (A)	His (H)	Lys (K)	Tyr (Y)	His (H)	STOP

A **missense mutation** is a **point mutation** (i.e. a single base substitution) that results in a codon that results in a different amino acid. In the above sequence, the first codon was changed from CUG to GUG changing the amino acid from Leucine to Valine.

d. Write a nucleotide sequence and translation to an amino acid sequence of a single base change producing a nonsense mutation.

…	CUG	GCC	CAC	UAG	UAU	CAC	UAA
…	Val (V)	Ala (A)	His (H)	STOP	Tyr (Y)	His (H)	STOP

A **nonsense mutation** is a **point mutation** (i.e. a single base substitution) that results premature stop codon. In the above sequence, the fourth codon was changed from AAG (Lysine) to UAG (a stop codon).

e. Write a nucleotide sequence and translation to an amino acid sequence of a single base change producing an extension of the protein.

…	CUG	GCC	CAC	AAG	UAU	CAC	UAU
…	Leu (L)	Ala (A)	His (H)	Lys (K)	Tyr (Y)	His (H)	Tyr (Y)

In this nucleotide sequence, the last nucleotide was changed from an A to a U resulting in the stop codon being changed to Tyrosine. This change will result in the protein continuing translation after what otherwise would have been the stop codon.

Problem #5

a. Copy the alignment obtained from the CLUSTAL Omega tool.

```
CLUSTAL O(1.2.1) multiple sequence alignment


gi|37222316|gb|AY350716.1|Lagothrix      GAGCAGCTGAACAAGCTGATGACCACCCTCCACAGCACTGCACCCCATTTTGTCCGCTGT
gi|37222318|gb|AY350717.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222320|gb|AY350718.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222322|gb|AY350719.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222324|gb|AY350720.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222326|gb|AY350721.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222328|gb|AY350722.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|45752610:36298-36383                  GAGCAGCTGAACAAGCTGATGACCACCCTCCATAG--CCGCACCCCATTTTGTCCGCTGT
                                         ******************************** **  * ********************

gi|37222316|gb|AY350716.1|Lagothrix      ATTGTGCCCAATGAGTTTAAGCAGTCAG
gi|37222318|gb|AY350717.1|               ATTGTCCCCAATGAGTTTAAGCAATCGG
gi|37222320|gb|AY350718.1|               ATTGTCCCCAATGAGTTTAAGCAATCGG
gi|37222322|gb|AY350719.1|               ATTATCCCCAATGAGTTTAAGCAATCGG
gi|37222324|gb|AY350720.1|               ATTATCCCCAATGAGTTTAAGCAATCGG
gi|37222326|gb|AY350721.1|               ATTATCCCCAATGAGTTTAAGCAATCGG
gi|37222328|gb|AY350722.1|               ATTATCCCCAATGAGTTTAAGCAATCGG
gi|45752610:36298-36383                  ATTATCCCCAATGAGTTTAAGCAATCGG
                                         *** * ***************** ** *
```

b. What kind of substitutions do we have? Count the number of transitions and transversions.

A **transition** occurs when there is a mutation from one purine to another purine (Adenine ↔ Guanine) or from one pyrimidine to another pyrimidine (Thymine ↔ Cytosine). A **transversion** is a swap from a purine to a pyrimidine (or vice versa). The output below is modified to mark transitions with an **I** and transversions with a **V**. Deletions are marked with a "D".

```
gi|37222316|gb|AY350716.1|Lagothrix      GAGCAGCTGAACAAGCTGATGACCACCCTCCACAGCACTGCACCCCATTTTGTCCGCTGT
gi|37222318|gb|AY350717.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222320|gb|AY350718.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222322|gb|AY350719.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222324|gb|AY350720.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222326|gb|AY350721.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222328|gb|AY350722.1|               GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|45752610:36298-36383                  GAGCAGCTGAACAAGCTGATGACCACCCTCCATAG--CCGCACCCCATTTTGTCCGCTGT
                                         ********************************I**DD*I********************

gi|37222316|gb|AY350716.1|Lagothrix      ATTGTGCCCAATGAGTTTAAGCAGTCAG
gi|37222318|gb|AY350717.1|               ATTGTCCCCAATGAGTTTAAGCAATCGG
gi|37222320|gb|AY350718.1|               ATTGTCCCCAATGAGTTTAAGCAATCGG
gi|37222322|gb|AY350719.1|               ATTATCCCCAATGAGTTTAAGCAATCGG
gi|37222324|gb|AY350720.1|               ATTATCCCCAATGAGTTTAAGCAATCGG
gi|37222326|gb|AY350721.1|               ATTATCCCCAATGAGTTTAAGCAATCGG
gi|37222328|gb|AY350722.1|               ATTATCCCCAATGAGTTTAAGCAATCGG
gi|45752610:36298-36383                  ATTATCCCCAATGAGTTTAAGCAATCGG
                                         ***I*V*****************I**I*
```

Of the nucleotides in the sequence, there were 5 transitions and 1 transversion. With the exclusion of one transition, all transitions and the transversion were only in the Lagothrix. The one remaining transition was in three organisms (Lagothrix).

c. What type of indels (insertion/deletion) do you see? What are the consequences of the indels?

There is a deletion in the human pseudogene of two bases. Since this is an exon and the deletion is not divisible by three, it would cause a frameshift mutation. This could make the gene non-functional (which would make sense since this is a pseudogene).

Problem #6

The winners of the Nobel Prize in Chemistry in 2013 were: Martin Karplus, Michael Levitt, and Arieh Warshel. The award was “for the development of multiscale models for complex chemical systems.”

The models developed by Karplus *et. al.* use both classical and quantum mechanical theory to describe both large chemical systems and reactions. In the quantum chemical model, the electrons and the atomic nuclei are the primary particles of interest. In contrast, classical models describe the behavior of atoms or groups of atoms. Karplus et. al.’s models use the quantum models to describe part of a system and link that part of the system to its surroundings using classical models. These models show how the two regions in the system interact in a physically meaningful way.

The work of Karplus et. al. served as the starting point of further theoretical research and more accurate derivative models. Their work has been applied to organic chemistry, biochemistry, heterogeneous catalysis, and theoretical calculation of the spectrum of molecules dissolved in a liquid.

Problem #7

Given two sequences, S and T:

S = GCTAGTCAGATCTGACGCTA
T = GATGGTCACATCTGCCGC

a. Construct a simple dot plot (window size = 1). Does your plot reveal any regions of similarity?

	G	C	T	A	G	T	C	A	G	A	T	C	T	G	A	C	G	C	T	A
G	*				*				*					*			*			
A				*				*		*					*					*
T			*			*					*		*						*	
G	*				*				*					*			*			
G	*				*				*					*			*			
T			*			*					*		*						*	
C		*					*					*				*		*		
A				*				*		*					*					*
C		*					*					*				*		*		
A				*				*		*					*					*
T			*			*					*		*						*	
C		*					*					*				*		*		
T			*			*					*		*						*	
G	*				*				*					*			*			
C		*					*					*				*		*		
C		*					*					*				*		*		
G	*				*				*					*			*			
C		*					*					*				*		*		

Without any filtering, there is a great deal of spurious noise in the dot plot. However, there is a discernible pattern along the diagonal of the sequence with differences in nucleotides: 2, 4, 9, and 15. These nucleotide differences do however break up the pattern in the sequence.

b. Construct a dot plot using a sliding window of size 4 with stringency value = 3. Does this plot reveal any regions of similarity between the two sequences?

	G	C	T	A	G	T	C	A	G	A	T	C	T	G	A	C	G	C	T	A
G									*					*						
A																				
T			*																	
G				*					*											
G					*															
T						*							*							
C							*													
A								*												
C									*											
A										*										
T						*					*									
C												*								
T													*							
G														*						
C															*					
C																				
G																				
C																				

There is a much clearer pattern along the diagonal. It begins in the third nucleotide (T in both sequences) and continues through the end of the sequence. It also better filters out the single nucleotide mutations that were visible in the previous dot plot. There are

still some spurious dots in the graph which means the graph could possibly benefit from a larger window and tighter stringency requirement.

c. Which dot plot is better and why?

For this nucleotide sequence, the second dot plot (i.e. part B) provided better results. It filtered out the spurious noise that made detection of the similar sequence more difficult. What is more, it was able to filter over those single base differences that break up the similar sequence (as observed in part A).