Making a DNA Vivid: Predicting the Protein Secondary Structure of a Prokaryotic DNA Sequence *ab initio*

Helen Chen[1], Hsien-Yi Yang[2], Subin Yun[3]
[1]**Department of Neurobiology, Physiology and Behavior, University of California, Davis**
[2]**Department of Molecular and Cellular Biology, University of California, Davis**
[3]**Department of Computer Science, University of California, Davis**

Introduction

Understanding the Concept from Previous Work to the Topic

DNA is the building block and storage of genetic information. Deoxyribonucleotide exists in a double-stranded form that suffices the Chargaff's rule[1], which is an identity of hydrogen bonding of nitrogenous bases to account for the base pairing between two deoxyribonucleotides. Adenine double-hydrogen bonds with thymine and cytosine triple-hydrogen bonds with guanine. Following the discovery of DNA structure by Rosalind Franklin[2], Francis Crick in 1958 stated the central dogma of molecular biology[3] that a DNA sequence is transcribed into an RNA sequence that is then translated to a protein. In eukaryotes, after alternative splicing, mature RNA is obtained without introns. Yet, in prokaryotes, a functional RNA strand is obtained without any splicing. Precedingly, a given RNA sequence is translated into a corresponding amino acid sequence based on the RNA codon table. The inter-relationship between DNA, RNA and protein, thus, serves an important tool to study biology and medicine.

Understanding the central dogma of molecular biology provides scientists useful information to study protein secondary structures from sequence, structure to function. For example, the study on murine coronavirus glycoproteins starts from the determination of the nucleotide and deduced amino acid sequences of complementary DNA clones to the observation of heptad repeat of hydrophobic residues that forms an alpha-helical structure[4]. An important association is derived by the mechanistic study on the process of virus entry that certain orientations of hydrophobic regions of helices account for the stabilization of interactions between virus and the cellular membrane from the host. The discovery is of significant value that understanding the virus-host communication not only uncovers the underlying mechanism and the functionality of alpha-helices in the transmembrane space but also renders the opportunity for scientists to develop treatment therapeutics. The study serves as an excellent example to signify the importance of the inter-dependence of DNA and protein secondary structure as well as understanding functional and structural biology to explicit disease prognosis.

Presentation to the Problem

Based on previous work and studies related to the central dogma of molecular biology, we are curious about if there is an approach to automate the process to transcribe and translate a given DNA sequence and provide a rudimentary analysis of its protein secondary structures to understand the function and the conformation of a given DNA sequence.

Primary structure gives the information about the arrangement and locations of amino acid residues in proteins. Secondary structure renders its biological and chemical significance. For example, transmembrane helix usually contains a loop and a helix. Within the helix, amphiphilic amino acids prefer to be at the interface between polar-non-polar environments[5]. The center of the helix has a certain iteration of hydrophobic residues. In addition, protein secondary structure also helps to facilitate the correct orientation of an enzymatic reaction, which is important for its downstream cellular signaling. It is obvious to state that DNA sequence and the secondary structure of proteins are interdependent. For example, mutations of DNA sequence may render a mis-folding of protein and therefore, negatively impacts downstream activities.

We are not only writing a gene finder program that transcribes and translates a DNA sequence into the corresponding RNA sequence and protein, but also going to focus on examining five types of secondary structures from a direct measurement of identifying repeats in the protein and also an indirect measurement of propensities of alpha-, 3-10, and pi-helix as well as beta-sheet and beta-turn with reference indices from Chou and Fasman algorithm[6] in order to provide a prediction of protein secondary structures *in silico*. The main purpose of the study is, therefore, to predict secondary structures from a given DNA sequence *ab initio*.

Introducing the KCNH2 protein test case

In addition to the test case from the prompt, we are going to use the DNA sequence of KCNH2 protein to test the accuracy of our algorithm in terms of its ability to transcribe and translate the DNA, and predict the chemical and secondary structure of a protein sequence. The KCNH2 gene belongs to a large family of genes that provide instructions for making potassium channels. These channels, which transport positively charged ions of potassium out of cells, play key roles in a cell's ability to generate and transmit electrical signals. Ion channels in the KCNH family (EAG, ERG and ELK) are important for nervous system function, cardiac physiology, and cancer biology[7]. These channels are made of four protein subunits that assemble to form a pore that spans the cell membrane, as they are transmembrane proteins. The purpose of reverse engineering the KCNH2 protein is served for confirming the design of our algorithm.

Method

Description of the Algorithm: Implementations, Assumptions and Limitations

Part 1: The central dogma of molecular biology

We utilize programming language C for the implementation of all *in silico* experimentations. Our algorithm reads an unknown DNA sequence as input from 5' to 3' end. A `FOPEN` function is used to read **only one** text file containing the input sequence and the maximum length of an input strand is set to be 10000 characters. The text file should **only** contain `ATCG` four characters. The input is case insensitive because we use a `for` loop to convert any lower-case letters to upper case letters. If a file is not found, an error message will be shown. We, then, do a reverse

complementary sequence to find out the other strand of the input DNA duplex. Based on the input DNA sequence, there will be six possible open reading frames because each amino acid is translated from three consecutive DNA letters and one DNA duplex has two directionalities, the leading strand and the lagging strand, making a total possibility of six. We define an open reading frame that starts from ATG (AUG in mRNA) which would be translated to methionine and stops once it reaches one of three possible stop codons, TAA, TGA and TAG (UAA, UGA, UAG in mRNA). An open reading frame must start from ATG and should only contain one stop codon. Again, we strictly define an open reading frame that starts at the first ATG and stops once it reaches the first stop codons no matter how many start and stop codons a sequence has. For example, **`ATGCCCTAAATGCCCCCCCCCTAA`** will output an open reading frame of **`ATGCCCTAA`**; it is not **`ATGCCCCCCCCCTAA`**. We compare six candidates and sort out the longest open reading frame based on our definition. Next, a transcription works by substituting thymine from the longest open reading frame with uracil. After the input DNA is processed and transcribed, each three of the RNA codons are translated into one corresponding amino acid. The directionality of translated peptide sequence is defined from the N terminus to the C terminus. The design of algorithm does not consider alternative splicing which is only occurring within the eukaryotic family. Therefore, the program is rendered to be used in the gene finder of a given prokaryotic DNA sequence, which contains only exons.

Part 2: Amino acid sequence property analysis

This section gives a rudimentary analysis of the output amino acid sequence based on the count and percentage of individual residue from the protein. In addition, four categories of amino acids are defined to sort out amino acid residues based on their chemical properties[8]. Non-polar residues are **`'I', 'A', 'F', 'L', 'M', 'P', 'V', 'W'`**; polar residues are **`'C','G','T','S','Y','Q','N'`**; charged residues are **`'H','K','D','E','R'`**, and aromatic residues are **`'F','W','Y','H'`**. Finally, a Grand Average of Hydropathy index (GRAVY) can be calculated as the sum of the hydropathy values for all the amino acids in a protein divided by the total number of residues in it. GRAVY follows the equation below.

$$GRAVY \approx \frac{1}{n}\sum_{k=1}^{k=n} x_k$$

n = number of residues, x is the Kyte-Doolittle hydropathy score[9] of residue k

Part 3: Protein secondary structure direct measurements

Based on Kyte-Doolittle hydrophobicity values[9] of residue positions, patterns of hydrophobic amino acid residues are observed to have certain iterations for different types of secondary structures. Alpha-helices have hydrophobic amino acids on n, n+4, n+7, and n+11 etc. positions in the sequence. 3-10 helices have hydrophobic amino acids on n, n+3, n+6, and n+9 etc. positions in the sequence. Pi-helices are unstable and rare, having positions of hydrophobic amino acids to appear in the sequence with a pattern of n, n+5, n+10, and n+15 etc. Similarly, beta-pleated sheets have hydrophobic amino acids on n, n + 2, n + 4, and n + 6 etc. positions in

the sequence. Beta-turns, on the other hand, follow an identity[10] of n, n+1, n+2, and n+3. Amino acid residues for n are `'N','C','D'`; for n+1 are `'P','S','K'`; for n+2 are `'N','D','G'`; for n+3 are `'W','G','Y'`. The position detector will output the starting position of the repeat and the ending position will be the last residue of a complete circle. For example, if an alpha-helical structure is detected at the position of 12; the repeat will end at the position of 23. The position detector gives the starting location of a found secondary structure within the peptide chain reading from the N-terminus to the C-terminus. We define that a direct measurement will be made only when a complete cycle of iteration of a protein secondary structure is detected. If no detection is found, it will not provide an absolute biological meaning; rather, it will be directed to indirect measurements to calculate its alpha-helical, beta-sheet and turn propensities.

Part 4: Protein secondary structure indirect measurements

An estimate of protein secondary structure propensity scale is designed to compensate for the difficulty of direct measurements of secondary structures in a given peptide sequence. We are using propensity indices from the Chou and Fasman algorithm[6]. In this project, we define a secondary structure that is predicted when it has the highest propensity compared to others as well as having a value of greater than one[11]. Based on the Chou and Fasman algorithm, A simplified equation is used below.

$$average\ propensity\ scale\ \approx \frac{1}{n}\sum_{k=1}^{k=n} x_k$$

n = number of residues, x is the propensity index of residue k from the Chou and Fasman algorithm

Brief analysis

For this project, we have tried to identify the longest open reading frame because, in biology, the translated protein sequence is usually the longest open reading frame. We designed our algorithm by reading the input from the first start codon to the first stop codon with a total possibility of six, since a DNA duplex has two directionalities in a replication fork. However, if we could implement this algorithm in a different way, we would like to improve the detection of the longest open reading frame in any given position of the input; not just limited to the first start and the first stop codon.

Results

Part 1: Test Case from the Prompt

Input:

`5'TCAATGTAACGCGCTACCCGGAGCTCTGGGCCCAAATTTCATCCACT3'`

Rudimentary Analysis

For the test case given in the project prompt, our program first generates its reverse complementary sequence `5'AGTGGATGAAATTTGGGCCCAGAGCTCCGGGTAGCGCGTTACATTGA3',` and then finds the longest possible "gene" from start codon `(ATG)` to stop codon `(TGA)` after determining its open reading frame

5'ATG,AAA,TTT,GGG,CCC,AGA,GCT,CCG,GGT,AGC,GCG,TTA,CAT,TGA3'. The corresponding mRNA sequence which substitutes T with U is also generated for translation: **5'AUGAAAUUUGGGCCCAGAGCUCCGGGUAGCGCGUUACAUUGA3′**. According to the genetic code chart, this mRNA sequence is translated to a protein with sequence **Nter-MKFGPRAPGSALH-Cter.** In this protein, there are one Leu, Phe, Met, Ser, His, Lys, and Arg each composed 7.69% of the protein and two Ala, Gly, and Pro each composes 15.38% of the protein. Among these residues, there are 53.85% of non-polar residues, 23.08% of polar residues and charged residues and 15.38% of aromatic residues. From these calculated percentages biologists can intuitively know the amino acid composition of this protein. Based on Kyte-Doolittle scale the GRAVY is -0.3308, indicating that this protein is hydrophilic.

Secondary Structure Prediction

Due to the simplicity of the original test case, we are not able to detect any pattern of hydrophobic amino acids in this sequence to directly measure its possible secondary structure, thus our algorithm shows **'No possible position of a secondary structure is found'** and **'No direct measurement of possible secondary structure is detected'.** However, our program is able to indirectly predict beta turn in the sequence, for the Chou and Fasman Score for β-turn is 1.0831 which is larger than 1, and the Chou and Fasman Scores for α-Helix and β-Sheet are smaller than 1.

Part 2: Test Case KCNH2 Protein

Input:

ATGGCGGCCCCAGCCGGGAAGGCGAGCAGGACAGGGGCTCTGCGGCCCAGGGCCCAGAAAGGCCGGGTGAGGCGGGCCGTGCGCATCTCCAGCCTCGTGGCCCAGGAGGTCCTGTCCCTGGGCGCCGA
CGTGCTGCCCGAGTACAAGCTGCAGGCACCGCGCATCCACCGCTGGACCATCCTGCATTACAGCCCCTTCAAGGCCGTGTGGGACTGGCTCATCCTGCTGCTGGTCATCTACACGGCTGTCTTCACAC
CCTACTCGGCTGCCTTCCTGCTGAAGGAGACGGAAGAAGGCCCGCCTGCTACCGAGTGTGGCTACGCCTGCCAGCCGCTGGCTGTGGTGGACCTCATCGTGGACATCATGTTCATTGTGGACATCCTC
ATCAACTTCCGCACCACCTACGTCAATGCCAACGAGGAGGTGGTCAGCCACCCCGGCCGCATCGCCGTCCACTACTTCAAGGGCTGGTTCCTCATCGACATGGTGGCCGCCATCCCCTTCGACCTGCT
CATCTTCGGCTCTGGCTCTGAGGAGCTGATCGGGCTGCTGAAGACTGCGCGGCTGCTGCGGCTGGTGCGCGTGGCGCGGAAGCTGGATCGCTACTCAGAGTACGGCGCGGCCGTGCTGTTCTTGCTCA
TGTGCACCTTTGCGCTCATCGCGCACTGGCTAGCCTGCATCTGGTACGCCATCGGCAACATGGAGCAGCCACACATGGACTCACGCATCGGCTGGCTGCACAACCTGGGCGACCAGATAGGCAAACCC
TACAACAGCAGCGGCCTGGGCGGCCCCTCCATCAAGGACAAGTATGTGACGGCGCTCTACTTCACCTTCAGCAGCCTCACCAGTGTGGGCTTCGGCAACGTCTCTCCCAACACCAACTCAGAGAAGAT
CTTCTCCATCTGCGTCATGCTCATTGGCTCCCTCATGTATGCTAGCATCTTCGGCAACGTGTCGGCCATCATCCAGCGGCTGTACTCGGGCACAGCCCGCTACCACACACAGATGCTGCGGGTGCGGG
AGTTCATCCGCTTCCACCAGATCCCCAATCCCCTGCGCCAGTGCTGTGACCACCCCGGGGCCTGGCCCCACTTCCACATCCCCGCTGTTGCCCGTCAGCCCCCTCCCCACCCTCACCTTGGACTCGCT
TTCTCAGGTTTCCCAGTTCATGGCGTGTGA

Rudimentary Analysis

In order to rectify our algorithm, we ran a second test case to enhance the reproducibility of the program. The reverse complementary sequence is as the following.

5'TCACACGCCATGAACTGGGAAACCTGAGAAAGCGAGTCCAAGGTGAGGGTGGGGAGGGGGCTGACGGGCAACAGCGGGGATGTGGAAGTGGGGCCAGGCCCCGGGGGTGGTCACAGCACTGGCGCAG
GGGATTGGGGATCTGGTGGAAGCGGATGAACTCCCGCACCCGCAGCATCTGTGTGTGGTAGCGGGCTGTGCCCGAGTACAGCCGCTGGATGATGGCCGACACGTTGCCGAAGATGCTAGCATACATGA
GGGAGCCAATGAGCATGACGCAGATGGAGAAGATCTTCTCTGAGTTGGTGTTGGGAGAGACGTTGCCGAAGCCCACACTGGTGAGGCTGCTGAAGGTGAAGTAGAGCGCCGTCACATACTTGTCCTTG
ATGGAGGGGCCGCCCAGGCCGCTGCTGTTGTAGGGTTTGCCTATCTGGTCGCCCAGGTTGTGCAGCCAGCCGATGCGTGAGTCCATGTGTGGCTGCTCCATGTTGCCGATGGCGTACCAGATGCAGGC
TAGCCAGTGCGCGATGAGCGCAAAGGTGCACATGAGCAAGAACAGCACGGCCGCGCCGTACTCTGAGTAGCGATCCAGCTTCCGCGCCACGCGCACCAGCCGCAGCAGCCGCGCAGTCTTCAGCAGCC
CGATCAGCTCCTCAGAGCCAGAGCCGAAGATGAGCAGGTCGAAGGGGATGGCGGCCACCATGTCGATGAGGAACCAGCCCTTGAAGTAGTGGACGGCGATGCGGCCGGGGTGGCTGACCACCTCCTCG
TTGGCATTGACGTAGGTGGTGCGGAAGTTGATGAGGATGTCCACAATGAACATGATGTCCACGATGAGGTCCACCACAGCCAGCGGCTGGCAGGCGTAGCCACACTCGGTAGCAGGCGGGCCTTCTTC
CGTCTCCTTCAGCAGGAAGGCAGCCGAGTAGGGTGTGAAGACAGCCGTGTAGATGACCAGCAGCAGGATGAGCCAGTCCCACACGGCCTTGAAGGGGCTGTAATGCAGGATGGTCCAGCGGTGGATGC
GCGGTGCCTGCAGCTTGTACTCGGGCAGCACGTCGGCGCCCAGGGACAGGACCTCCTGGGCCACGAGGCTGGAGATGCGCACGGCCCGCCTCACCCGGCCTTTCTGGGCCCTGGGCCGCAGAGCCCCT
GTCCTGCTCGCCTTCCCGGCTGGGGCCGCCAT3'

As expected, it finds the original input to be the longest possible gene out of six possible open reading frames from start codon ATG to stop codon TGA.

5'AUGGCGGCCCCAGCCGGGAAGGCGAGCAGGACAGGGGCUCUGCGGCCCAGGGCCCAGAAAGGCCGGGUGAGGCGGGCCGUGCGCAUCUCCAGCCUCGUGGCCCAGGAGGUCCUGUCCCUGGGCGCC
GACGUGCUGCCCGAGUACAAGCUGCAGGCACCGCGCAUCCACCGCUGGACCAUCCUGCAUUACAGCCCCUUCAAGGCCGUGUGGGACUGGCUCAUCCUGCUGCUGGUCAUCUACACGGCUGUCUUCAC
ACCCUACUCGGCUGCCUUCCUGCUGAAGGAGACGGAAGAAGGCCCGCCUGCUACCGAGUGUGGCUACGCCUGCCAGCCGCUGGCUGUGGUGGACCUCAUCGUGGACAUCAUGUUCAUUGUGGACAUCC
UCAUCAACUUCCGCACCACCUACGUCAAUGCCAACGAGGAGGUGGUCAGCCACCCCGGCCGCAUCGCCGUCCACUACUUCAAGGGCUGGUUCCUCAUCGACAUGGUGGCCGCCAUCCCCUUCGACCUG
CUCAUCUUCGGCUCUGGCUCUGAGGAGCUGAUCGGGCUGCUGAAGACUGCGCGGCUGCUGCGGCUGGUGCGCGUGGCGCGGAAGCUGGAUCGCUACUCAGAGUACGGCGCGGCCGUGCUGUUCUUGCU
CAUGUGCACCUUUGCGCUCAUCGCGCACUGGCUAGCCUGCAUCUGGUACGCCAUCGGCAACAUGGAGCAGCCACACAUGGACUCACGCAUCGGCUGGCUGCACAACCUGGGCGACCAGAUAGGCAAAC
CCUACAACAGCAGCGGCCUGGGCGGCCCCUCCAUCAAGGACAAGUAUGUGACGGCGCUCUACUUCACCUUCAGCAGCCUCACCAGUGUGGGCUUCGGCAACGUCUCUCCCAACACCAACUCAGAGAAG
AUCUUCUCCAUCUGCGUCAUGCUCAUUGGCUCCCUCAUGUAUGCUAGCAUCUUCGGCAACGUGUCGGCCAUCAUCCAGCGGCUGUACUCGGGCACAGCCCGCUACCACACACAGAUGCUGCGGGUGCG
GGAGUUCAUCCGCUUCCACCAGAUCCCCAAUCCCCUGCGCCAGUGCUGUGACCACCCCGGGGCCUGGCCCCACUUCCACAUCCCCGCUGUUGCCCGUCAGCCCCCUCCCCACCCUCACCUUGGACUCG
CUUUCUCAGGUUUCCCAGUUCAUGGCGUGUGA3'

After transcribing input DNA sequence to mRNA sequence , our program generates the translated protein sequence **Nter-**
MAAPAGKASRTGALRPRAQKGRVRRAVRISSLVAQEVLSLGADVLPEYKLQAPRIHRWTILHYSPFKAVWDWLI
LLLVIYTAVFTPYSAAFLLKETEEGPPATECGYACQPLAVVDLIVDIMFIVDILINFRTTYVNANEEVVSHPGR
IAVHYFKGWFLIDMVAAIPFDLLIFGSGSEELIGLLKTARLLRLVRVARKLDRYSEYGAAVLFLLMCTFALIAH
WLACIWYAIGNMEQPHMDSRIGWLHNLGDQIGKPYNSSGLGGPSIKDKYVTALYFTFSSLTSVGFGNVSPNTNS
EKIFSICVMLIGSLMYASIFGNVSAIIQRLYSGTARYHTQMLRVREFIRFHQIPNPLRQCCDHPGAWPHFHIPA
VARQPPPHPHLGLAFSGFPVHGV-Cter. While KCNH2 protein is made up of all of the 20 amino acids, the six most abundant ones are Leu (10.94%), Ala (9.41%), Ile (7.63%), Val (7.38%), Pro (6.36%) and Ser (6.36%) according to 'Amino Acid Calculator' of the program. It is noticeable that 'Residue Property Calculator' of the program calculates the percentage of nonpolar amino acids in this sequence to be 51.4%, meaning non-polar amino acids make up more than half of the total KCNH2 amino acids which matches the following results.

Secondary Structure Prediction

Our program detects 1 possible 3-10 helical structure starting at the 118th amino acid (Ile) and 1 pi-helical structure starting at the 3rd amino acid (Ala) in output KCNH2 protein sequence as direct measurements of various secondary structures based of the pattern of appearance of hydrophobic amino acids. Through indirect measurements of secondary structures the Chou and Fasman Scores of alpha-helix and beta-sheet are both larger than 1 (1.01 and 1.0305); because the propensity of beta-sheet is greater than that of alpha-helix and both of them are greater than 1, "**a possible beta-sheet" is predicted**".

Discussion

The gene finder program generates a translated protein sequence from a given DNA sequence. The central dogma of molecular biology has different variants in eukaryotic and prokaryotic family. In this project, we only consider the first start and stop codon to find out the longest open reading frame. For future research, we would like to search any start and stop codons in a given sequence in any direction to enhance the usability and applicability of the program. Throughout the project, we only consider the prokaryotic family that the transcribed RNA does not contain introns. The degeneracy of genetic codons gives the freedom for silent

mutations, rendering the tolerance of a change in DNA codons may not result in a loss of protein function.

The amino acid sequence property analysis gives a big picture of protein residue composition and its biochemical properties. Percent specific categories of amino acids not only shows the composition but also correlates to the hydrophobicity of the protein, since a hydrophobic protein contains a majority of non-polar residues to avoid contact with water. The grand average hydrophobicity (GRAVY) gives an overall hydrophobicity propensity estimate of the protein primary structure. Positive GRAVY values indicate that the protein is hydrophobic, and negative values means that the protein is hydrophilic. GRAVY not only functions for scientists to understand the environment of the protein but also provides clues to unravel protein functions.

The secondary structure prediction gives the understanding of both direct measurement and indirect measurement. A direct measurement provides the starting position of a found secondary structure. We focus on examining protein secondary structures of helices, sheets and turns. Helices can be further broken down into alpha-, 3-10- and pi-helices characterized by different $\Phi \; and \; \psi$ angles. Alpha helix is the most dominant form followed by 3-10- and pi-helices. Also, if a direct measurement of a protein secondary structure cannot be found, an indirect measurement will be made based on the Chou and Fasman algorithm that is believed to have a 55-60% success rate of predictions[12]. Nevertheless, a direct measurement only gives a possible detection of a secondary structure; it is not a guaranteed search-and-match function because a random coil may have the same repeat identity without orienting itself to the preferred configuration.

Understanding the homolog of the *ab initio* predicted protein secondary structure is important to uncover its phylogenetic relationships as well as its molecular functions. BLAST is used by biologists to find a homolog for a sequence. Here we use BLASTp which is a protein to protein BLAST to test the accuracy of transcription and translation of our program. We blast our output KCNH2 protein sequence from the program as the query to find out if the output has the same homologs in a genetic database as the pre-existing protein sequence. The result from **Figure 2** shows that our reverse-engineered protein sequence has a query cover 100% with KCNH2 protein. Also, the E value indicates that the probability due to chance, that there is another alignment with a similarity greater than the given score is zero, showing an exact match has been found. The BLASTp result of KCNH2 rectifies the design of our algorithm for a successful translation from an input DNA sequence.

On the one hand, the test case from the prompt has a turn propensity with a negative GRAVY, since most beta-turns function to change directions of a protein sequence. In addition, the majority of beta-turn residues are composed of hydrophilic residues, so it is not surprising to have a negative GRAVY score from the analysis. On the other hand, it is observed that KCNH2 matches the biological fact that non-polar (hydrophobic) amino acids are found to be the most abundant in the protein embedded in a lipid bilayer, as potassium channel KCNH2 protein is a transmembrane protein which contains six transmembrane segments in each of its four subunits,

and it functions to conduct flows of potassium ions out of the cell when voltage varies. The detection from our direct measurement of 3-10- and pi-helical structures gives a possible estimate of helical structure sitting within the channel to facilitate ion transport. The positive GRAVY score, moreover, signifies the overall hydrophobic propensity of the transmembrane space of a channel protein. As for the confirmation of reproducibility, we examine the reverse-engineered sequence of KCNH2 by generating the homolog modeling via Swiss-Model. From **Figure 3**, The top-hit model shows a density map of possible beta-sheets and helices observed from the Ramachandran plot with the density of helices higher than that of beta-sheets. In addition, from **Figure 4**, the computationally-build homolog model clearly indicates helical structures as well as a few beta-sheets, thus solidifying our algorithm.

KCNH2 protein [Homo sapiens]

Sequence ID: AAI27674.1 **Length:** 393 **Number of Matches:** 1

Range 1: 1 to 393 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
810 bits(2092)	0.0	Compositional matrix adjust.	393/393(100%)	393/393(100%)	0/393(0%)

```
Query  1    MAAPAGKASRTGALRPRAQKGRVRRAVRISSLVAQEVLSLGADVLPEYKLQAPRIHRWTI  60
            MAAPAGKASRTGALRPRAQKGRVRRAVRISSLVAQEVLSLGADVLPEYKLQAPRIHRWTI
Sbjct  1    MAAPAGKASRTGALRPRAQKGRVRRAVRISSLVAQEVLSLGADVLPEYKLQAPRIHRWTI  60

Query  61   LHYSPFKAVWDWLILLLVIYTAVFTPYSAAFLLKETEEGPPATECGYACQPLAVVDLIVD  120
            LHYSPFKAVWDWLILLLVIYTAVFTPYSAAFLLKETEEGPPATECGYACQPLAVVDLIVD
Sbjct  61   LHYSPFKAVWDWLILLLVIYTAVFTPYSAAFLLKETEEGPPATECGYACQPLAVVDLIVD  120

Query  121  IMFIVDILINFRTTYVNANEEVVSHPGRIAVHYFKGWFLIDMVAAIPFDLLIFGSGSEEL  180
            IMFIVDILINFRTTYVNANEEVVSHPGRIAVHYFKGWFLIDMVAAIPFDLLIFGSGSEEL
Sbjct  121  IMFIVDILINFRTTYVNANEEVVSHPGRIAVHYFKGWFLIDMVAAIPFDLLIFGSGSEEL  180

Query  181  IGLLKTARLLRLVRVARKLDRYSEYGAAVLFLLMCTFALIAHWLACIWYAIGNMEQPHMD  240
            IGLLKTARLLRLVRVARKLDRYSEYGAAVLFLLMCTFALIAHWLACIWYAIGNMEQPHMD
Sbjct  181  IGLLKTARLLRLVRVARKLDRYSEYGAAVLFLLMCTFALIAHWLACIWYAIGNMEQPHMD  240

Query  241  SRIGWLHNLGDQIGKPYNSSGLGGPSIKDKYVTALYFTFSSLTSVGFGNVSPNTNSEKIF  300
            SRIGWLHNLGDQIGKPYNSSGLGGPSIKDKYVTALYFTFSSLTSVGFGNVSPNTNSEKIF
Sbjct  241  SRIGWLHNLGDQIGKPYNSSGLGGPSIKDKYVTALYFTFSSLTSVGFGNVSPNTNSEKIF  300

Query  301  SICVMLIGSLMYASIFGNVSAIIQRLYSGTARYHTQMLRVREFIRFHQIPNPLRQCCDHP  360
            SICVMLIGSLMYASIFGNVSAIIQRLYSGTARYHTQMLRVREFIRFHQIPNPLRQCCDHP
Sbjct  301  SICVMLIGSLMYASIFGNVSAIIQRLYSGTARYHTQMLRVREFIRFHQIPNPLRQCCDHP  360

Query  361  GAWPHFHIPAVARQPPPHPHLGLAFSGFPVHGV  393
            GAWPHFHIPAVARQPPPHPHLGLAFSGFPVHGV
Sbjct  361  GAWPHFHIPAVARQPPPHPHLGLAFSGFPVHGV  393
```

Figure 2

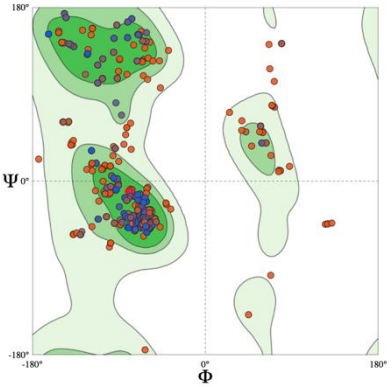


Figure 3

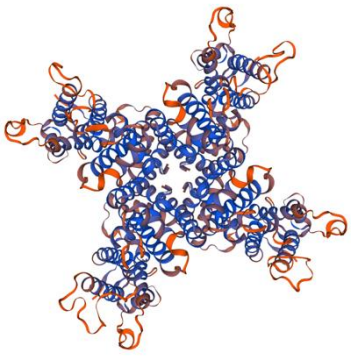


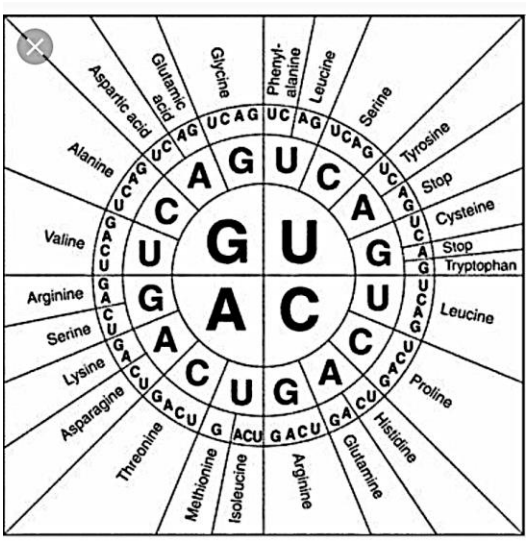
Figure 4

Bibliography

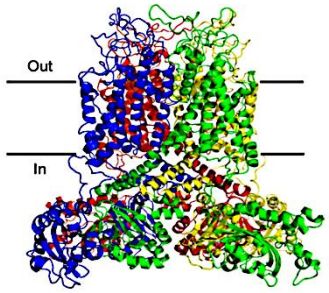
1. Forsdyke, D. R., & Mortimer, J. R. (2000). Chargaff's legacy. *Gene*, *261*(1), 127-137.
2. Klug, A. (1968). Rosalind Franklin and the discovery of the structure of DNA. *Nature*, *219*(5156), 808-810.
3. Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561-563.
4. Chambers, P., Pringle, C. R., & Easton, A. J. (1990). Heptad repeat sequences are located adjacent to hydrophobic regions in several types of virus fusion glycoproteins. *Journal of General Virology*, *71*(12), 3075-3080.
5. Mitaku, S., Hirokawa, T., & Tsuji, T. (2002). Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane–water interfaces. *Bioinformatics*, *18*(4), 608-616.
6. Chou, P. Y., & Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annual review of biochemistry*, *47*(1), 251-276.
7. Duff H.J. (2013) hERG (KCNH2) Potassium Channel, Function, Structure and Implications for Health and Disease. In: Kretsinger R.H., Uversky V.N., Permyakov E.A. (eds) *Encyclopedia of Metalloproteins*. Springer, New York, NY
8. Zhu, C., Gao, Y., Li, H., Meng, S., Li, L., Francisco, J. S., & Zeng, X. C. (2016). Characterizing hydrophobicity of amino acid side chains in a protein environment via measuring contact angle of a water nanodroplet on planar peptide network. *Proceedings of the National Academy of Sciences*, *113*(46), 12946-12951.
9. Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, *157*(1), 105-132.
10. Chou, P. Y., & Fasman, G. D. (1977). β-Turns in proteins. *Journal of molecular biology*, *115*(2), 135-175.
11. Chen, H., Gu, F., & Huang, Z. (2006). Improved Chou-Fasman method for protein secondary structure prediction. *BMC bioinformatics*, *7*(4), S14.
12. Chou, P. Y., & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, *13*(2), 222-245.

Appendix

RNA codon table



KCNH2 from the literature[7]



Chou and Fasman Propensity Scale

P_α			P_β			P_t		f_i		f_{i+1}		f_{i+2}		f_{i+3}	
Glu	1.51	H_α	Val	1.70	H_β	Asn	1.56	Asn	0.161	Pro	0.301	Asn	0.191	Trp	0.167
Met	1.45		Ile	1.60		Gly	1.56	Cys	0.149	Ser	0.139	Gly	0.190	Gly	0.152
Ala	1.42		Tyr	1.47		Pro	1.52	Asp	0.147	Lys	0.115	Asp	0.179	Cys	0.128
Leu	1.21		Phe	1.38	h_β	Asp	1.46	His	0.140	Asp	0.110	Ser	0.125	Tyr	0.125
Lys	1.16	h_α	Trp	1.37		Ser	1.43	Ser	0.120	Thr	0.108	Cys	0.117	Ser	0.106
Phe	1.13		Leu	1.30		Cys	1.19	Pro	0.102	Arg	0.106	Tyr	0.114	Gln	0.098
Gln	1.11		Cys	1.19		Tyr	1.14	Gly	0.102	Gln	0.098	Arg	0.099	Lys	0.095
Trp	1.08		Thr	1.19		Lys	1.01	Thr	0.086	Gly	0.085	His	0.093	Asn	0.091
Ile	1.08		Gln	1.10		Gln	0.98	Tyr	0.082	Asn	0.083	Glu	0.077	Arg	0.085
Val	1.06		Met	1.05		Thr	0.96	Trp	0.077	Met	0.082	Lys	0.072	Asp	0.081
Asp	1.01	I_α	Arg	0.93	i_β	Trp	0.96	Gln	0.074	Ala	0.076	Thr	0.065	Thr	0.079
His	1.00		Asn	0.89		Arg	0.95	Arg	0.070	Tyr	0.065	Phe	0.065	Leu	0.070
Arg	0.98	i_α	His	0.87		His	0.95	Met	0.068	Glu	0.060	Trp	0.064	Pro	0.068
Thr	0.83		Ala	0.83		Glu	0.74	Val	0.062	Cys	0.053	Gln	0.037	Phe	0.065
Ser	0.77		Ser	0.75	b_β	Ala	0.66	Leu	0.061	Val	0.048	Leu	0.036	Glu	0.064
Cys	0.70		Gly	0.75		Met	0.60	Ala	0.060	His	0.047	Ala	0.035	Ala	0.058
Tyr	0.69	b_α	Lys	0.74		Phe	0.60	Phe	0.059	Phe	0.041	Pro	0.034	Ile	0.056
Asn	0.67		Pro	0.55	B_β	Leu	0.59	Glu	0.056	Ile	0.034	Val	0.028	Met	0.055
Pro	0.57	B_α	Asp	0.54		Val	0.50	Lys	0.055	Leu	0.025	Met	0.014	His	0.054
Gly	0.57		Glu	0.37		Ile	0.47	Ile	0.043	Trp	0.013	Ile	0.013	Val	0.053

Hydropathy scale and information used in the assignments

Side-chain	Hydropathy index	$\Delta G^{\circ}_{transfer}$ (water–vapor)[a]	Fraction of side-chains 100% buried[b]	Fraction of side-chains 95% buried[c]
Isoleucine	4·5	4·4	4·5	5·2
Valine	4·2	4·2	4·3	4·2
Leucine	3·8	4·5	3·2	2·8
Phenylalanine	2·8	2·5	2·5	3·5
Cysteine/cystine	2·5	1·9	6·0	3·2
Methionine	1·9	1·9	1·0	1·9
Alanine	1·8	3·9	5·3	1·6
Glycine	−0·4	—	4·2	1·3
Threonine	−0·7	−0·6	−0·5	−1·0
Tryptophan	−0·9	−0·9	−2·4	−0·3
Serine	−0·8	−0·8	−0·7	−1·0
Tyrosine	−1·3	−1·1	−3·3	−2·2
Proline	−1·6	—	−2·4	−1·8
Histidine	−3·2	−4·2	−3·6	−1·9
Glutamic acid	−3·5	−3·9	−2·8	−1·7
Glutamine	−3·5	−3·5	−4·0	−3·6
Aspartic acid	−3·5	−4·5	−2·5	−2·3
Asparagine	−3·5	−3·8	−3·1	−2·7
Lysine	−3·9	−3·2	—	−4·2
Arginine	−4·5	—	—	—