Drew Hotson

Prof. Brutlag

Bioinformatics 118

March 13, 2000

Mutations to the CFTR protein

Striking approximately one in 2000 Caucasians, cystic fibrosis (CF) is the most common lethal genetic disorder among the group. The leading cause of death for CF patients is respiratory failure due to lung infections in the thick mucus of the lung. The disease effects the epithelial cells that line the lungs and digestive tract, causing thick mucus and pancreatic cysts. Specifically, the cells have a failed cystic fibrosis transmembrane conductance regulator (CFTR) protein, which functions as a chloride channel. Patients with CF are unable to transport chloride ions through their epithelial cell membranes.

J.R. Riordan and colleagues in Toronto first cloned the CFTR gene in 1989. The nucleotide sequence was translated into the primary amino acid sequence to reveal a protein with 1480 amino acids. The amino acid sequence was found to be homologous to nucleotide (ATP)-binding folds (NBFs) in two places. Additionally, there were two sets of six hydrophobic regions capable of spanning the lipid bilayer of the cell membrane, thus embedding the protein in the membrane. Lastly, between the two NBFs was a highly charged regulatory (R) domain that showed similarity to sequences for phosphorylation by protein kinases A or C (PKA and PKC). This information was used to construct the possible protein structure for CFTR. Because of the amphiphatic nature of CFTR, the protein can't be crystallized for the exact structure to be determined by X-

ray crystallography, so the amino acid sequence provides the best determination of the protein structure.

Both the R domain and the ATP binding sites are believed to regulate the chloride channel. The R domain is composed of alternating clusters of positively and negatively charged amino acids, with 69 of the 241 residues being polar. Cyclic AMP stimulates PKA phosphorylation of serine residues in the R domain in an ATP reducing reaction. Not until the R domain is phosphotylated can the NBFs bind ATP. The R domain is coded for by exon 13, which is residues 590 to 831.

The NBFs of CFTR occur on both halves of the protein in three highly conserved segments. The first half is residues 433 to 473, 488 to 513, and 542 to 584, while the second half is residues 1219 to 1259, 1277 to 1302, and 1340 to 1382. Once the R domain has been phoshporylated, the NBFs can bind ATP. Reduction of ATP induces a conformational change in the protein, which opens a channel between the transmembrane sequences that allows chloride to pass out of the cell.

The necessity for conservation of the NBF regions can be demonstrated by the computer program eBLOCKS. This program creates blocks of amino acid sequences from proteins from the same family, and compares a given sequence to the blocks to determine if it would likely fit the motif of the blocks. For the CFTR sequence, 541 blocks were returned, and the first 200 were examined, all with a specificity less than $8.0e^{-15}$. Forty of the first 200 blocks were for CFTR, and the remaining 160 were for various other ATP-binding proteins, such as ATP-binding cassette transporter ABC, multidrug resistance-associated protein 1, and ATP-dependent bile acid permease. Of the 160 blocks for these ATP-binding proteins, 61 were matched between residues 433 to

584 of CFTR, and 62 between residues 1219 and 1382. This verifies that those regions are the NBFs, and suggests mutations in those areas may be detrimental because they are so highly conserved among ATP-binding proteins.

The most common patient mutation to CFTR is ΔF508, a deletion of the phenylalanine in the 508th amino acid position, which accounts for around 70% of the cases of cystic fibrosis. This specific mutation is disproportionately high because heterozygotes are better adapted to survive typhoid fever, so the allele increased in the European population just as the sickle cell anemia allele is unusually common in Africa for the resistance to malaria offered to heterozygotes. Still, there are many other known CFTR mutations, and these are the ones most commonly found in other ethnicities.

The known information about the protein structure can be used to help predict where in the protein CF causing mutations may appear. It is difficult to correlate the severity of the phenotype with the genotype, because there are so few people with rare mutations, and it is difficult to assign a severity based on only one or two cases. However, there is no denying that many mutations are less severe than ΔF508. Some patients do not discover until adulthood that their respiratory or digestive disease is actually due to a mutated CFTR. Others show only the symptom of congenital bilateral aplasia of the vas deferens (CBAVD) for their CFTR mutation.

A CFTR mutation database has been established on the internet. It lists all known mutations to the protein, and how they are believed to be caused. While it can't specify the exact severity associated with a given mutation, it does list which mutations are believed to cause only CBAVD. To find what specific parts of the protein are most important, the mutations that effect large portions of the protein are ignored. For

example, no consideration was paid to frameshifts, mutations into stop codons, mutations causing promoter or splicing errors, or large deletions or insertions. Rather, single amino acid substitutions, insertions, and deletions were examined to find the exact area of the protein that is important for functionality.

Using the protein structure, it is predicted that mutations in the NBFs and R domain will have the greatest effect. This is because if the R domain does not get phosphorylated, or the NBFs don't bind ATP, the chloride channel cannot open. Mutations in the transmembrane region may not be quite as severe. These regions are created by hydrophobic alpha-helices, so severe loss of functionalities are likely caused by changing an amino acid from hydrophobic to hydrophilic, deletion of a hydrophobic amino acid, or insertion of a hydrophilic amino acid.

Analysis of the 404 single amino acid substitutions, insertions, and deletions from the CFTR mutation database shows that mutations are actually most likely to occur in the transmembrane regions as well as the first NBF, and least likely in the ends of the protein and the R domain. The ends of the protein are defined as the first 49 amino acids and the 98 amino acids after the second NBF. The transmembrane regions are from residues 50 to 432 and residues 832 to 1218. The NBFs are residues 433 to 584 and 1219 to 1382, while the R domain is from residue 590 to 831. To determine the likelihood of a mutation, the ratio of the percentage mutations on a given region compared to the percentage of amino acids of the protein on the region was calculated. A number greater than 1 means there is more likely to be a mutation on the region, because there is a larger percentage of mutations than amino acids. For example, on the first transmembrane

region, which is made up of 25.9% of the amino acids, 29.2% of the mutations occur. This gives a likelihood number of 1.12.

One underlying assumption is made. As rare mutations to CFTR are discovered only when a patient shows some phenotype and is screened for a mutation, those mutations that cause no change in protein function go undiscovered. If it is assumed that there is an equal probability of each amino acid being altered, but that only the ones that cause a loss of functionality are reported, than the regions with the most reported mutations are the most important for functionality.

The most telling information is that the R domain had a likelihood number much less than 1. It is known that when the R domain, which is regulated by PKA, is partially deleted the chloride channel is active with ATP even without PKA. This suggests that mutations to the R domain often don't hinder phosphorylation, but rather prevent the R domain from serving as a regulator of the chloride channel, still allowing functionality. Another possibility is that only some of the amino acids in the R domain are required to for phosphorylation to occur. Fourteen of the 30 mutations involve changing a charged amino acid, which make up only 28% of the region. This suggests that charged amino acids may be important for phosphorylation, while others aren't.

It is also interesting that the transmembrane regions are likely to have mutations. It was predicted that the leading cause of mutation in this region would be due to substitution of a hydrophilic amino acid for a hydrophobic one. Indeed, this was the case for 40 of the 118 mutations in the first region, or 34%, when statistically it would occur less than 25% if it were random. Yet there are still 78 mutations that aren't of this

variety. They may disrupt the alpha-helix, or alter the conformation of the protein in some other way that obstructs functionality.

It is odd that as predicted the first NBF had a high likelihood of being mutated, but the second was neutral. This may indicate that the likelihood of mutation may not depend on the function of a region, because two similar regions have different likelihoods. However, it is more likely due to statistical error from the small sample size, because the other homologous regions, being the transmembrane regions and the start and end of the protein, both have similar likelihoods of mutation. What the CFTR mutation database doesn't fully explain is the severity of the disease, though it is likely more severe for mutations in these regions. The most common mutation, ΔF508, is in the first NBF, and gives patients the standard, severe CF. If a mutation in these regions prevents the NBF from binding ATP, the chloride channel cannot function at all, and the phenotype will be severe.

It is not surprising that the ends of the protein are least likely to have mutations. They have no role in the protein other than conformation, so mutations that are not severe enough to alter the conformation should have little effect. This is likely the case, because both the start and end of the protein have likelihood numbers less than 1.

As far as genotype to phenotype relations, the CFTR mutation database does report which mutations are believed to cause only CBAVD. These mutations were only found in the transmembrane and NBF regions, and with a similar frequency. Nine out of 115 mutations in the NBFs, or 7.8%, and 15 out of 252 transmembrane mutations, or 6.0%, caused only CBAVD. The mutations in the NBF likely allow some degree of channel activity, and probably can bind ATP at a reduced rate. The transmembrane

mutations also are not as severe, possibly because they allow the chloride channel to open, just with greater difficulty. Five of the 15 transmembrane mutations involved the substitution of a hydrophilic amino acid for a hydrophobic one, which is the same ratio as for CF causing mutations. It is difficult to distinguish which CFTR mutations cause CF and which cause only CBAVD, and one can only speculate as to why a given genotype determines the severity of the phenotype.

Analysis of the CFTR mutation database shows the NBFs and transmembrane regions to be the most susceptible to causing loss of functionality due to mutation. The R domain and ends of the protein are both better able to withstand being mutated while still showing normal phenotype. This suggests that the NBFs and transmembrane regions are most important for the chloride channel to work. The R domain and ends of the protein are less important for functionality.

Appendix

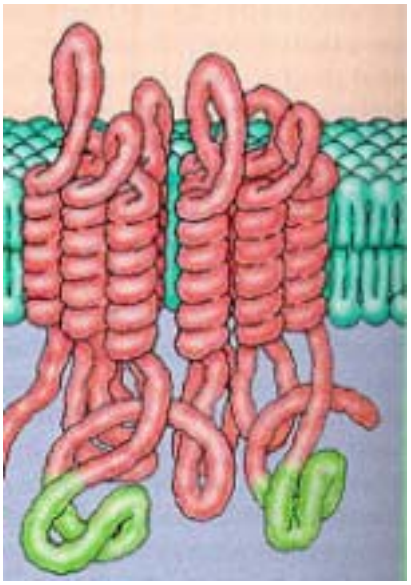

Figure 1 (From Cooper). The CFTR protein is made of two sets of six transmembrane regions, two nucleotide binding folds (in green), and the R domain between the NBFs.

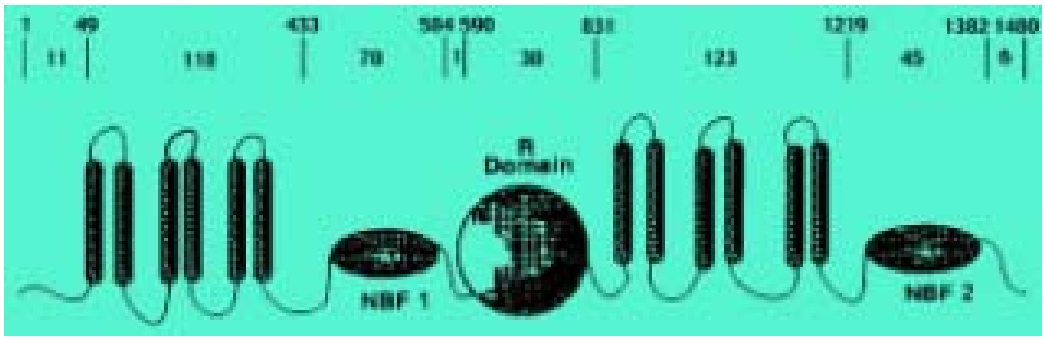

Figure 2 (From Collins). This diagrams the CFTR protein stretched out. The numbers above the hash marks are the residue numbers, and the numbers between the hashes represent the number of mutations in the region.

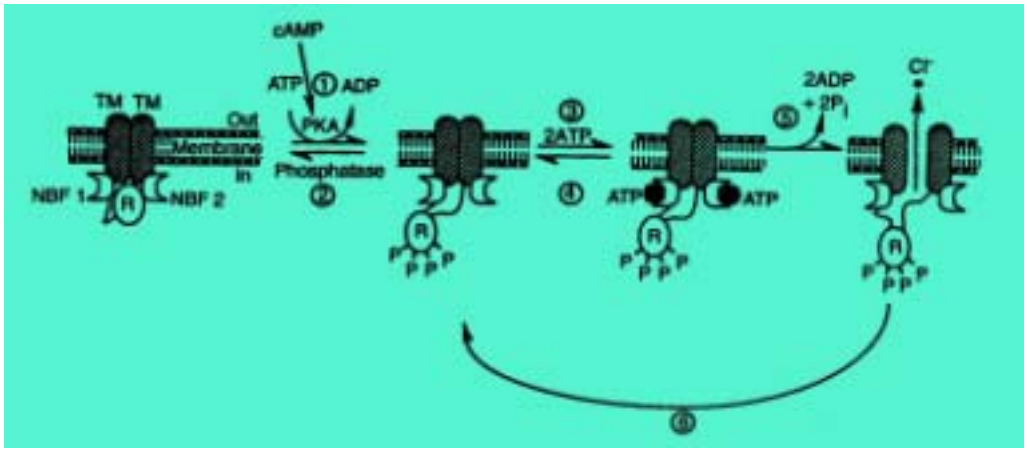

Figure 3 (From Collins). This shows the pathway to open the chloride channel. First the R domain is phosphorylated by PKA in an ATP reducing reaction. Then the NBFs bind and reduce ATP, opening the channel.

Region	Start	Trans-membrane 1	NBF 1	R domain	Trans-membrane 2	NBF 2	End
% protein	3.3	25.9	10.2	16.3	26.2	11.0	6.6
% mutation	2.7	29.2	17.3	7.4	30.4	11.1	1.5
Likelihood	0.82	1.12	1.69	0.45	1.16	1.01	0.23

Table 1. This table shows what percentage of the amino acids are in a given region, what percentage of the mutations are in the region, and the likelihood of a mutation occurring in the region. Likelihoods greater than one mean it is likely to have a mutation, while less than 1 means it is unlikely.

References

Collins, Francis. S. “Cystic fibrosis: molecular biology and therapeutic implications.” *Science* 256: 774-779.

Cooper, Geoffrey M. The Cell: A Molecular Approach. Washington D.C.: The American Society for Microbiology Press, 1997.

McKusick, Victor A. “*602421 Cystic Fibrosis Transmembrane Conductance Regulator; CFTR.” Online Mendelian Inheritance in Man, March 7, 1998. http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?602421.

Riordan, J. R.; Rommens, J. M.; Kerem, B.; Alon, N.; Rozmahel, R.; Grzelczak, Z.; Zielenski, J.; Lok, S.; Plavsic, N.; Chou, J. L.; Drumm, M. L.; Iannuzzi, M. C.; Collins, F. S.; Tsui, L.-C. “Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA.” *Science* 245, 1989: 1066-1073.

Su, Qiaojuan; Brutlag, Douglas. “eBlocks.” Stanford University: Brutlag Bioinformatics Group, September 13, 1999. http://eblocks.stanford.edu.

Tsui, Lap-Chee. “CFTR Mutation Table.” http://www.genet.sickkids.on.ca/cftr-cgi-bin/FullTable.