The number of human genes and proteins

Daniel B. Davison
Applied Genomics, Bristol-Myers Squibb Pharmaceutical Research Institute, 311
Pennington-Rocky Hill Road, Pennington, NJ 08534 USA. ddavison@bms.com

Abstract

The definitive prototype for nanotechnology is the cell. Its many machines and exquisitely controlled internal and external movements are a reference for all researchers working in the field. The complete instructions for every molecular machine in a cell is specified in its DNA. The interactions of those parts are emergent properties of the individual components (RNAs, fats, sugars, and proteins). At present, there is considerable controversy among biologists regarding the number of human genes and proteins. In part, the differences stem from differences in definition. In this presentation we will define a gene as a transcription unit. Each transcription unit may have zero to many splice forms (known as "alternative splices"). While there are several methods for gene prediction, all involving computational tools, none agree. Estimates range from 20,000 to 120,000. One way to approach this question, involving both computation and experiment, is to look at copies of fragments of messenger RNA (mRNA), called expressed sequence tags (ESTs). mRNA comes only from a gene being expressed by a cell or tissue. By clustering mRNA fragments, we can try to reconstruct the expressed gene. The final result is a very rough representation of the 'true expressed transcript'. Our results consistently demonstrate that there are some 70,000 transcription units with an average of 1.2 different transcripts per transcription unit. Thus, we estimate the total number of human genes at about 85,000. Post-translational modification will make the total number of proteins be much higher.

Keywords: EST clustering, number of human genes, genomics

Introduction

There are already self-replicating nanomachines that work under very high temperatures and pressures while others work at standard temperature and pressures. Each nanomachine consists of a large number of smaller machines, most of which self-assemble. In biological terms, the ribosome, the transcriptional apparatus, and the replication apparatus all represent elegantly subtle solutions to a wide range of physical problems. Inside the cell, there are poorly understood mechanisms that allow proteins to hitch a ride on the cell's internal scaffold (the cytoskeleton) and move from one place to another. In the case of your big toe, a protein synthesized in your spinal cord is precisely delivered to the nerve ending in your toe – a distance of about a meter! A landmark in human knowledge was recently reached with the announcement of the completion of the draft sequence of the human genome. This represents the opening of a completely new phase in understanding what it is to be human, in the factors that influence many human diseases, and in understanding nature's solutions to the problems discussed in this volume and at this conference. One key datum of great interest is the number of human genes and proteins. While the question is simple enough to state, in practice it is very hard to answer. Not the least is the issue of carefully defining what one means by the term "gene". Before doing that, we'll first present background information, and then present our estimate of the number of human genes and proteins.

A brief introduction to molecular biology

The central dogma of molecular biology states that information moves from DNA to RNA to protein. Every cell has all of its instructions encoded in its deoxyribonucleic acid (DNA). DNA is made up of four subunits, or bases, termed A, C, G, and T for adenine, cytosine, guanine, and thymine, respectively. These subunits are connected together into very long chains, called chromosomes. The vast majority of species have linear chromosomes. Most species have a different number of pairs of chromosomes, even those that are evolutionarily closely related. Humans have 23 pairs of chromosomes, 22 autosomes and two sex-determining chromosomes, the X and the Y. They range in size from 50 to 263 million base pairs (see http://alces.med.umn.edu/tables/humchr.html). The total size of the human genome is about 3 billion bases.

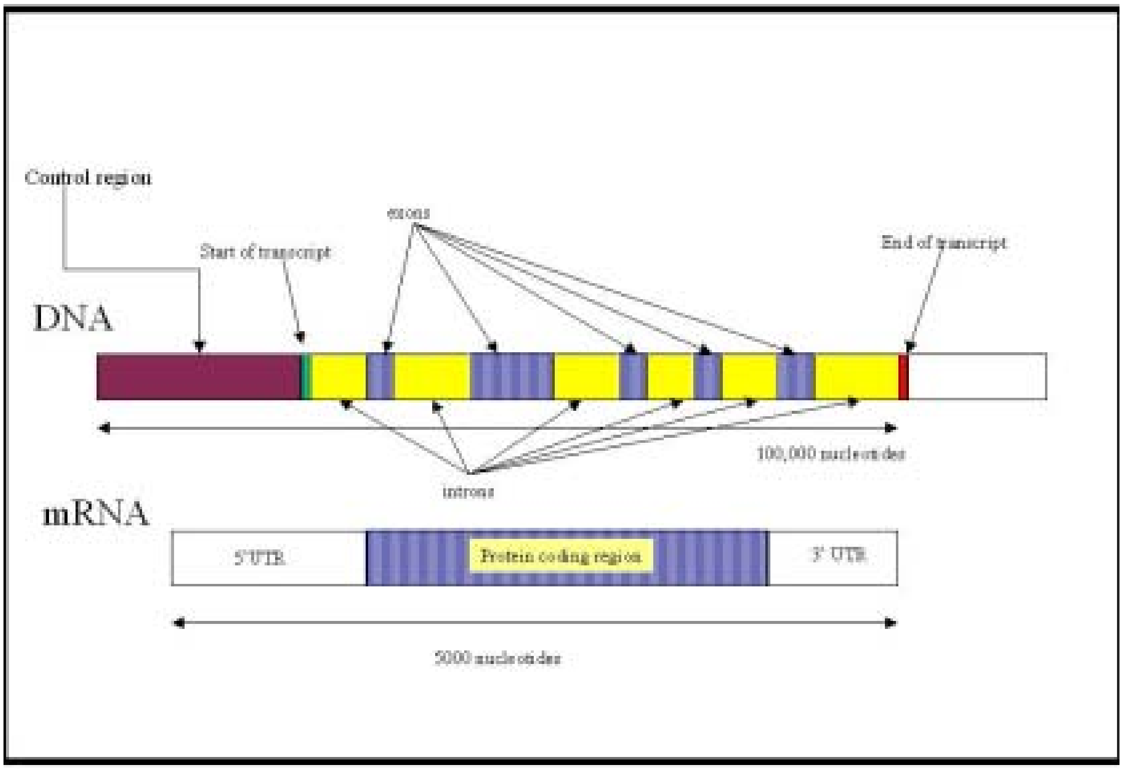

Figure 1. A schematic view of some features of a eukaryotic transcription unit. The top line represents the DNA and the bottom, the messenger RNA. The region between “start of transcript” to “end of transcript” is copied into mRNA, then the internal non-coding regions (introns) are removed, leaving the protein coding regions (exons) together with some upstream sequence (the 5’ untranslated region [UTR]) and downstream sequence (3’ UTR).

A gene has a number of characteristics, diagrammed in Figure 1. It is the basic unit of heredity, that which is passed from generation to generation. It usually encodes a protein. The gene is copied into an mRNA in a process called “transcription”. That mRNA is then translated into a protein sequence by a very large protein-RNA molecular machine called the ribosome. Each gene has a non-transcribed region upstream and downstream of the coding region involved in regulation its expression. Some parts that are transcribed into mRNA have parts that are cut out afterward (introns) leaving behind the instructions for making the protein (the exons). Furthermore, while every cell has the genes for making every protein, not all genes are expressed (translated into protein) in all cells. Some genes have a role in only one kind of tissue (e.g. the lung) and would not be expressed in another tissue (e.g. the skin).

Each of these features of the gene, introns, exons, and regulatory regions, are still relatively poorly understood. The signals that distinguish one from the other are subtle. Therefore, at present the best way to locate genes is experimentally. Researchers isolate cells of interest and extract the total RNA, a process called “library construction”. The mRNAs are separated from the structural RNAs, then copied into DNA with a special enzyme. These pieces of DNA are called complementary DNA, cDNA. During this process the RNA is frequently partially degraded. The cDNA fragments are called expressed sequence tags, or ESTs. A particular gene may not be expressed in a particular cell type, so it would not be present in the library. Other genes may be very highly expressed, so there would be many ESTs from those genes. All cells have certain metabolic functions and structural features that they must have to live, so those "housekeeping" ESTs will be found in all cells.

By examining all the ESTs available from cells in a wide variety of tissues and developmental states, one can get an idea of how many expressed genes there are. Another complication is that a gene may produce more than one transcript by including or excluding specific exons or by altering the length of a specific exon. This is known as "alternative splicing".

EST analysis

A common way to put ESTs together is by cluster analysis. The goal of such a project is the construction of a gene index, where ESTs and full-length transcripts are partitioned into index classes (or clusters) such that they are put into the same index class if and only if they represent the same gene. Accurate gene indexing facilitates gene expression studies, and reduces the cost of gene discovery. Also, effective gene clustering serves as a starting point for the discovery of new gene expression variants such as alternative splicing forms. Torney et al [1] have developed an algorithm, called "d^2" that is used as the basis for a program we have developed termed d^2_cluster. It is an agglomerative algorithm specifically developed for rapidly and accurately partitioning transcript databases into index classes by clustering ESTs and full-length sequences according to minimal linkage or "transitive closure" rules.

The d^2 algorithm

In 1989, Torney et al.[1] presented an algorithm called d^2, for comparing two sequences. Most sequence comparison algorithms are context-dependent. By this, we mean that one can obtain a traditional sequence alignment (Figure 2). A set of letters from one sequence can be written over a set of letters from another sequence and lines drawn between related or identical letters. In other domains, this is called approximate string matching. In contrast, a context-independent, word-based method such as d-squared asks only if the substrings (words) of a particular size occur the same number of times in both sequences, regardless of location (Figure 3).

Figure 2. Sequence Alignment. A portion of a FASTA[2] comparison of the bacterial mobile DNA element IS1 from *Escherichia coli* with a distantly related element, IS1νξ (isois1), from *Shigella dysenteriae.*

```
>>INS1ECLAC                                               (884 nt)
 initn: 472 init1: 472 opt: 641
 55.656% identity in 778 nt overlap

                            10        20        30        40
isois1              GCATCGATATTTTTTCAGGTGATGCCTCTAATTAGTTGAATCTGATG
                                     ::::::::  : :: :   :: :: :  ::
INS1EC CTGATAAGAGACACCGGCATACTCTGCGACGGTGATGCTGCCAACTTACTG-ATTTAGTG
          30        40        50        60        70         80
```

Figure 3. Idealized d^2 sequence comparison. The core idea is that the same words occur the same number of times in both sequence s. The blocks of color denote similar sequences as detected by d^2. The blocks are not in linear order between Sequence 1 and Seq uence 2, thus, no alignment, as shown in Figure 2, is possible.

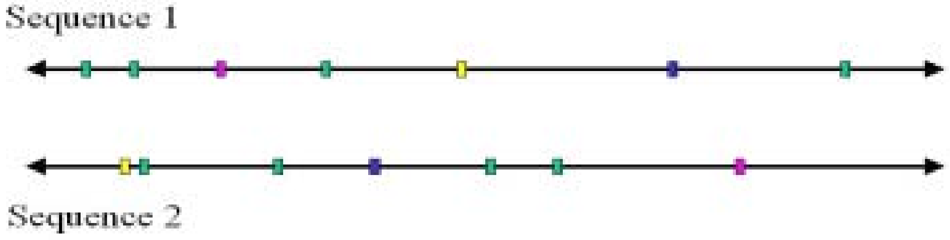

Table 1. Determination of the average number of alternative splices per cluster

Stringency	Subclusters	Clusters	Ratio	Singletons
80%	82,657	67,499	1.2	149,510
85%	82,113	66,654	1.2	141,345
90%	82,907	67,410	1.2	145,263
99%	81,858	70,852	1.2	234,436

Clustering Public ESTs

Turning to the results of clustering human ESTs [3], we used ESTs available in the public domain as of February 11, 2000. There were 1,373,183 ESTs in the input file. The DoubleTwist, Inc. (http://www.doubletwist.com) Clustering and Alignment Tools, version 3.5, were used to screen, cluster, and assemble the data on a 4 processors of an 8-processor Silicon Graphics Origin 2000 with 3GB of RAM. The standard screening files supplied by DoubleTwist were used. Standard input parameters were used except for the stringency of d^2 comparison, the variable called *set_d2_string.* We used values of 0.8, 0.85, 0.9, and 0.99, corresponding to an identity of 80%, 85%, 90%, and 99%, respectively. These values caused the program to be increasingly stringent in its clustering. Sequences must have at least *set_d2_string* identity to be joined into a cluster, as described above. Table 1 summarizes the results of these clusterings.

At a stringency of 80%, there are 82,657 subclusters (of any type) in 67,499 clusters, or 1.2 transcripts per cluster. This result places a loose lower bound on the estimate of the number of genes for this stringency. There must be about this many *transcripts* in the data set. If we assume each cluster is a unique transcript, this stringency implies there are about 67,500 parent transcripts. However, this is an underestimate for several reasons. First, the public data do not reflect transcripts from all possible tissue types, disease types, and developmental states. Additionally, technical details of the generation of ESTs limit how many ESTs can be obtained from a particular tissue. Very rare transcripts, for example, those occurring less than 10 times in a tissue, are unlikely to be represented. Therefore, estimating the number of genes, transcripts, and proteins involves some guesswork as to the level of underrepresented genes and transcripts in the publicly available data. We will not attempt to make an estimate of the number of genes included in this category. So, from the data given above, we estimate that there is an average of 1.2 transcripts per gene, giving a total of some 81,000 different transcripts in the human genome. This means that there are at least 81,000 different proteins, before post-translational modification.

The Number of Human Genes and Proteins

Based on the data presented above, we believe there are between 1.2 and 1.5 transcripts per transcription unit in the human genome. Our clustering suggests to us that there are about 70,000 transcription units in the human genome. This implies that there are up to 84,000 different proteins produced in a human cell, before post-translational modification, such as the attachment of phosphate, adenylate, lipids, or sugar groups. This is obviously an extremely coarse estimate.

Acknowledgments

The author would like to thank David Torney, Winston Hide, and John Burke for their essential roles in the development of d2 and d2_cluster.

References

1. Torney, D.C., et al. “Computation of d2: A Measure of Sequence Dissimilarity”
Computers and DNA, SFI Studies in the Sciences of Complexity, vol. VII, eds. G. Bell and T. Marr, Addison-Wesley. (1990).

2. Pearson WR. “Flexible sequence similarity searching with the FASTA3 program package” Methods Mol Biol. 132:185-219 (2000).

3. Burke J, Davison D, Hide W. “d2_cluster: a validated method for clustering EST and full-length cDNA sequences” Genome Res. Nov;9(11):1135-42. (1998)