ALBERTO RIVA: Alberto Riva, I'm an instructor at CHB. I'm going to talk to you today about the most important resources for finding and using biomedical information, especially information connected with the study of the human genome.

So this is going to be something probably slightly different from what you've heard so far, which concentrate more on actual locations, and most of them will be websites, where you can find information, and to talk about how this information is stored and represented, how it's accessible, and what it can be used for.

So you're going to see a long list of references to sites, websites, with URLs. Don't worry if you can't remember all of them because, of course, I'm going to distribute the slides and will be easier to just look them up.

So I'm going to start with something that you probably heard about many times before. The so-called central dogma of molecular biology, as you know, almost all our cells contain DNA in their nucleus-- DNA is the molecule that encodes information-- for at least from the purposes of this presentation, this is what we're interested in.

And this information is transcribed into our name molecules, that then exit the nucleus under the form of mRNA. MRNA is then translated into proteins. And proteins are ultimately what is responsible for essentially all the external manifestations, all the observable properties of our biology.

So I'm talking about things like metabolism, general physiology, diversity between individuals, diseases, drug response. These are all, n some way or another, due to the different proteins that act within our cells and outside our cells.

So we have names, of course, for the extremes of the spectrum. We call genotype-- this information is encoded in the DNA. And on the other hand, we call phenotype whatever is at the other extreme of the spectrum, anything we can observe, we can measure, from the outside.

So what I'm going to try to show you is that, as you move from one end of the spectrum to the other, you're going to encounter very different forms of information, of data. And each one has its own specific nature and function, and needs to be treated with different tools. And needs to be represented in different ways.

So essentially these are the questions that I'm going to answer. How is all this information represented? What are the different ways that we can store and describe this information? Where does it come from, where is it stored, how do we find retrieve and use it?

So we've talked about the two ends of the spectrum. So there are some very deep differences between the kind of information you find when you're talking about the genotype, and when we're talking about phenotypes. For example, the genotype is digital, because each base pair in our DNA can be exactly represented using one of four symbols, A, P, G, C. Then you can also think about insertions, deletions, and so on.

But essentially, using a small number of symbols, you can provide an exact representation of our genome, of all the 3 billion base pairs that compose our DNA.

On the other hand, the phenotype is, say, analog. Because most phenotypes are qualitative in nature. They cannot be measured exactly or precise, they cannot even be defined precisely in most cases. You always have to take into account the effect of environmental factors that, again, are very hard to describe in a quantitative way.

At the root of all this is that one of the biggest problems in the study of protein is the fact that the proteins are not uniquely determined by their sequence. For DNA, you just look at the sequence and you know essentially all that there is to know about DNA.

For proteins, you cannot look at the subsequence of a protein and understand just by looking at it what the protein is going to do. Not even how it's going to be-- not even what it's three-dimensional structure is going to be. That's difficult enough. Then understanding what the protein does, just by looking at the sequence, is still very far from being feasible.

On the other hand, it's interesting to see that our knowledge of these things has progressed in the opposite direction. Because obviously, it's much easier to observe a phenotype than to observe DNA in ourselves. So the first thought is of inherited traits date back to Mendel in 1866. And DNA was discovered, more or less, the same years. But at the time, nobody had any idea that there was any connection between these two things, between the DNA and inherited traits.

It took over 80 years for this concept to be proven. So the definite proof that genes are made of DNA dates back to 1952. After that, progress was faster because the elucidation of the structure of DNA, and the DNA replication mechanism, came one year later.

Then the geniculate code was deciphered between 1961 and 1966. Something that we now take for granted, like the discovery of introns, only happened in 1977. And finally, the Human Genome Project, that was officially declared a success last year, brought us to the point where we now know the exact base pair sequence of our genome.

So we now know with a sufficient degree of certainty-- we're going to talk about this more later-- we know the exact base pair composition of the human genome. And also several other genomes, but of course, the human one is the one we're most interested in.

Going back again from genotype to phenotype, there is another thing to note. I've just said that we now have the complete sequence for our genome, but of course, this is an approximation. It's an abstraction, actually, because even if we're all human beings, there are no two human beings that are exactly the same. And this is a consequence of the fact that there are differences between the DNA of two any human beings.

These differences are due to polymorphisms, like single nucleotide polymorphisms, so locations which, instead of having the base that everybody else has, you have a different base. Microsatellites, repeats, insertions, deletions, translocations, these are all things that can happen to your DNA sequence that can modify in ways that, of course, are not enough to turn you into another animal-- you're still a human being-- but your DNA sequence is slightly different from the sequence of any other human being.

So the other is, one of these polymorphisms, every 1,000 bases. So if you think we have 3 billion base pairs, it adds up to a very large number of differences. Which means that when you study the human genome, I mean, we now have the sequence of the human genome. But then if you go and look at one individual, you're not going to find that his DNA matches exactly the sequence that you find in the human genome databases.

You're going to find approximately one difference every 1,000 bases. And understanding what these differences do and mean, and what is their consequence, is one of the most interesting problems in current bioinformatics and molecular biology, because now we finally have the tools of looking at our genome with this level of detail. We can look at individual base pairs and we can see, well, there should be an A here, and instead, we have a C. Does that cause a problem?

So again, we're going to go back to this soon. And the same thing happens for phenotypes, although in a slightly different ways. Phenotypes are generalizations, too. So when we talk about things like species.

Again, I said, we're all human beings, but we're all different from a geniculate point of view. So it means that putting all of us together into one big group, in one species, is a generalization, of course. And even going down to ethnicity, or even a concept like disease, these are generalizations, because these are concepts that cannot be defined in a precise, formal way.

So we will see as we go forward that we're going to encounter very different forms of data. And we're going to need different methods to manipulate this data, according to what the purpose of our work is. So just to make this clear, if we are working at the level of DNA, then the typical operations we might be interested in doing are, for example, sequence matching. So to understand if certain stretch of a sequence matches anything else that you've seen before.

So this is useful, for example, when you discover a new gene, you want to know, first of all, if it's really a new gene, or if it's already been seen somewhere else. And if it's a new gene, you would like to understand-- have an idea of what it does. And if you find a similarity between your new gene and something that's already known, that can give a lot of information.

We're talking about discovering genes-- finding genes in a DNA sequence is not trivial. There are programs that do this. They just look at the sequence and they find locations in the DNA sequence that might contain genes. And I'm not going to go into details on this, but there are various reasons why this is a pretty complex thing to do from a computational point of view.

Homology searches-- again, these refer to looking for similarities between DNA sequences in different organisms. So if you discover the function of a certain gene in the mouse, for example, you would like to if it does the same thing in humans. So again, if you find a high degree of similarity between the two genes, you can hypothesize that they are also going to have the same function.

We've talked about polymorphisms, so another pretty common cooperation is performing DNA sequences, is SNP detection. So if you sequence the same stretch of DNA from a certain number of different individuals, then you can compare them. And you're going to find that most of the locations are the same for all individuals. But some of them will be different.

And this is how SNPs are discovered, how single nucleotide polymorphisms are discovered. They are locations where different individuals don't have the exact same nucleotide. And we'll talk later about why genotyping is important, what kind of information you can get from that, and how this relates to diseases, to basically trying to figure out what is the relationship between the genotype, and in this case, between polymorphisms in your genotype and phenotype, like disease.

At a level of RNA, it might be interesting to look at alternative splicing transcription rearrangements-- these are all things that happen to the original DNA sequence when it's transcribed into RNA. It undergoes a series of transformations that can, of course, affect, in a very deep way, the final product.

And this process, the process of transcribing the DNA sequences into RNA, is, of course, at the basis of expression analysis. And you've heard a lot in the other lectures about microarrays. So I'm not going to go into too much details on this, but differentiable analysis, clustering, and so on, these are all the usual things that can be done using microarrays, using gene expression microarrays.

We talk about proteins-- if you're studying a protein, the interesting thing is to do with a protein are trying to predict its active domains, in order to have an idea of how the protein might behave, what function it might have, how it might interact with other proteins, with other genes, and so on.

Predicting the three dimensional structure of a protein is another very important and very complex task. Starting homology and conservation of proteins across different organisms can give you a very good idea of the importance of some proteins. So if something that's been around for millions of years, it probably means that it's involved in a very basic mechanism, while there are some proteins that are new, that are only specific to human beings. And that, again, can give you some information.

And finally, something that is very challenging, and it's receiving a lot of attention lately, is the automatic construction and analysis of metabolic pathways and regulatory pathways. If you are able to understand how proteins interact with each other, and interact with the rest of the cell, how they then regulate other genes, and in turn other proteins, then you can use this information to try to build, in a computational way, the kind of pathway maps that biologists have been drawing by hand for decades.

And of course, we're still very far away from being able to do this in the general case. It works in some limited case, and we're going to see later some examples of some of these things. But these are all very challenging problems that, of course, are still very much open.

And finally, we get to the phenotype then, we could put a very long list of things here, but could talk about population genetics, about association studies. Association studies are studies that try to correlate the presence of a certain genotype with an observed phenotype. Like in the most common case, association studies based on SNPs, they just look at two different alleles of a SNP, and they try to figure out if there is statistical correlation between one or two alleles and the disease, and that might mean that the SNP is indeed responsible for the disease in some way.

And clinical trials, of course, to validate all this.

OK, so two more slides about philosophy and then we'll start with the more practical stuff.

I've already mentioned the word gene a lot of times, and I'm going to mention it again very often. So it might be interesting to ask ourselves, what is a gene? And this is something that it's a question to which the answer is probably obvious, but it turns out that there are actually many possible answers, according to the context you're in, according to the different view of the world that you're using.

So for example, if you ask a classical geneticist what is a gene, you will get the answer that a gene is the smallest unit of inheritance. This is the definition that goes back to Mendel, essentially. If you ask someone who is doing medical research, you will get the answer that the gene is a disease-causing trait. So hear about the [INAUDIBLE] gene, or the gene for cystic fibrosis, and so on.

So in this case, the word gene has a very clear clinical connotation. If you ask a molecular biologist, you get the answer that the gene is a recipe-- is essentially a program to build one or more proteins. And we can go on, we can ask biochemists, and you will get the answer that the gene is an element in a metabolic network. It's an active element in one of those big networks of interacting genes that regulate each other, and that overall realize the metabolic process.

If you ask a motor geneticist, you will get the answer that the gene is a locus on a chromosome, in a certain region of a chromosome, that has a functional characterization. Locus that was studied, and was found to have a specific function in our biology.

And finally, if you ask a bioinformatician, you will get the answer that a gene is just a stretch of DNA where we know there is a gene, because the database tells us that there is a gene there. So we know it has a transcription start site, encoding sequence start site, has exons and intrinsic in certain positions.

So in the following, we are to see examples of all of these different ways to look at the gene.

And to start, of course, we're going to start from the beginning. So from DNA sequence data. And now, we're starting to look at where all these different pieces of information I've told you about can be found, and how they're stored and represented.

So if we're talking about DNA sequence data, the first place to go is, of course, GenBank. GenBank is the largest repository of sequence data. It accepts direct submissions from researchers. So anybody in the world who sequences a new piece of DNA can send it to GenBank, and it's put in the big cauldron.

This is data from-- the most recent data that I could find from one year ago. It contained more than 22 million sequences and 100,000 distinct organisms. With a total of almost 30 billion nucleotides.

And this is the URL for GenBank. And GenBank is at the basis of the NCBI cluster. So the National Center for Biotechnology Information, is a branch of the NIH, that has the task of assembling the largest possible number of databases of biomedical information.

They manage GenBank, and GenBank, in turn, is at the basis for a lot of other resources that we're going to see now, that are all part of this cluster of NCBI resources. They're all interconnected, so you can easily jump from one to the other, and that is a very powerful way of exploring this kind of data.

So this is a graph that shows you the growth of GenBank in recent years. You can see the very steep growth of the number of base pairs. And the almost equally steep rate growth of the number of sequences. And you can probably tell that the number-- we're sequencing longer and longer sequences, because the blue graph grows more rapidly than the red one.

But anyway, what do you do when you have all these accumulated sequences? So in GenBank, you just have sequences by themselves that can be very short, very long, but they're just independent sequences that were put there by investigators. So the thing you can do, if you have enough sequences from the same organism, you can try assembling them, putting them together, and trying to reconstruct the entire genome. And this is what was done to assemble the human genome, for example, and all the other genomes that are being sequenced.

You start with-- you look at the sequences you have, and if you can find overlaps, then you know that these two sequences are related in some way, and you proceed from there. So let's say you've sequenced this sequence, then another one, they are distinct and they have an overlap, so you don't know where they are on the genome. But then, if you sequence a third one that overlaps the first one and includes the second one, then you know that you can basically arrange them in this way.

And then let's say you sequence another one, and again you find an overlap with one that you already have. So like this, and in the end you're going to build a map that tells you where all these fragments should be positioned on the chromosome. And you have different levels of coverage, so you have regions that you've seen only once, regions you've seen twice, regions you've seen three times.

If you have a sufficient degree of coverage, then you can say, well, OK, I believe this is the right sequence. And you proceed from there, you take that for granted, you proceed from there by connecting more and more overlapping pieces. And again, this is how the human genome was sequenced essentially.

Up to a level of coverage-- I think it's five or six times covered, so each stretch of DNA in the human genome has been sequenced at least five or six times for validation.

So and the details of how this process has been implemented and CVI are here.

And in addition to the human genome, of course, we have a lot of other genomes that are completed, or near completion. These numbers are probably higher by now we have over 1,000 viruses. And then many other organisms from different domains of life. Of course, eukaryotes are the hardest organisms to sequence.

But the human genome is considered finished by now. It's hard to go above this level of accuracy. It's probably not even necessary because the differences between two human beings are probably the same order of magnitude.

But now we have several other organisms, including a chimp, that was recently released. And it's going to be very interesting because it's essentially identical to the human genome. So there is 1%-- the differences are about 1% between the genomes of the human and the chimp. And it will be very interesting to see exactly where these differences are, and why these differences are responsible for the fact that we are humans and they are monkeys.

Again this is a link to the entry point for the description of all the genomes that are stored in NCBI. And a new genomes, the small ones-- the viruses and bacteria, there are new genomes, essentially, every week. So these numbers constantly change.

Now, we now have the complete subsequence of the human genome. Where do you find it? So the best resource, in my opinion, for looking at the human genome in literally is Golden Path. Golden Path is a genome browser for several different organisms. Initially it was only for human-- now it has mouse, rat, chimp, Drosophila, yeast, and a few others.

The nice thing about Golden Path is that it gives a graphical view, so you can see it in the next slide, it's very clear, it's very easy to find all the information you need about a certain region of the chromosome. On the other hand, all the information it provides is available in easy to download and to parse formats. So if you want to build your own database, it contains the same information that's something that is pretty easy.

It provides arbitrary DNA sequences-- so you can ask for any region of any human chromosome, you'll get back the exact DNA sequence for that region. For something that might think it's easy by now, but two or three years ago, it was still something that was almost impossible to obtain. And Golden Path was the first site to provide something like this.

It gives you the absolute position of all the known elements of our genome. So genes, markers, mutations, other features, they tell you exactly they are at this location, in terms of the absolute base pair.

This is the URL for Golden Path. And this is how it looks like-- this is an example, we're looking at the region that contains the TLR1 gene. And you can see all these different tracks that provide different information on different objects. So for example, up here we have the genes, we have different sets of known genes. We have predicted genes, according to different prediction algorithms.

We have the mRNAs that were aligned to this gene. We have a tract like this and that shows the conservation between human and mouse in this region. So you can see it's very interesting to see that obviously the coding part of the gene is the one that is most highly conserved between human and mouse-- that is the functional part. So it undergoes selective pressure.

There are tracks that tell you the location of SNPs, and so on. There are many others that of course don't fit in here, but you can customize the display. You can select the tracks you want to see, and you get your own view of a certain genetic region. You have the coordinates up here, chromosome 4, the band, and so on.

And this is just to show that you can query it for any-- this is the same DNA region we were looking at before. But in this case, we asked for the DNA sequence, and we get it.

There's another way of looking at the genome using the NCBI map viewer. It's essentially the same thing-- it's a graphical browser to look at genomes and annotations of genomes. It's organized around several maps-- they have sequence maps, cytogenetic maps, language maps, radiation hybrid, human mouse homology maps. So each one of these is a view that gives you a different set of objects in the view.

So in the sequence map, you can find information about the genes, transcripts, gene clusters, and so on. Inside the genetic map, for example, you find information about disease genes, what bands break points. It's extremely detailed, because of course, it can rely on the whole set of NCBI databases, so basically everything that could possibly want to know is in there.

I, personally, find it a bit complex to use, a bit harder to use than Golden Path. But it's is a matter of taste. It's organized in a different way.

Shows you much more detailed information here in this graphical bars, on the features of the genetic region. But then the coded information is a bit harder. They use all these abbreviations here.

So it's a matter of taste. They essentially serve the same purpose with different levels of detail in different areas.

Now, we've talked about SNPs. It's the only form of mutation we're going to talk about, but it's also the most important one. Because first of all, SNPs are the most common form of variation in our genome. They're much more frequent than microsatellites or insertions, deletions, and other things.

And they're important because for example, they can be used as genomic markers. So SNPs are at a fixed location in the genome. And if you know where the SNP is, you can find it-- you can find the same location in different individuals. So you can use them as markers. You can use them as causal candidates for diseases, because a certain percentage of the SNPs introduce changes that then have some consequence on the genotype, on the phenotype.

What I mean is that, for example, if you have a SNP in the coding subsequence of a protein, you're going to get a protein that has an abnormal sequence. And that can be a change that doesn't cause any consequence, or it might be a very dramatic change. For example, the most extreme example is there are some SNPs that introduce a stop in the protein sequence.

So the protein sequence is truncated-- instead of just being modified, it's truncated. It's shorter than it should be. As you can imagine, this is a chain that can be very dangerous. There are many diseases that are due to the fact that you have SNPs that truncate proteins.

They can be used as evolutionary markers, because SNPs arise randomly during the replication, and then they are transmitted from one generation to the next. And it's very interesting to study how SNPs get-- how the frequency of the SNP changes in a population.

So if you have a SNP that provides an advantage for you-- because I mean, most SNPs are deleterious. But in some cases, the SNP can also provide an advantage, if it generates something that was not present before, and that works better than the original. So if you have a SNP that introduces a change is beneficial, then you will-- given enough time-- you will see that the frequency of the SNPs increases in the population, more and more individuals are going to have the variant form of the SNP.

On the other hand, if a SNP is neutral, then there is no selective pressure, and it will either go away by chance, or will stay at a certain basic level of frequency. So you can study the frequency of the SNP to understand if it's undergoing selective pressure, so to know if it's deleterious or not, or you can use it to reconstruct basically the history of our genome. There are ways of calculating the age of the SNP, so when that mutation arose in the history of our genome.

Now the largest database of SNPs that we have again, is at NCBI, it's called the dbSNP-- currently contains over 4 million human SNPs-- actually, I think that by now this number is closer to 5 million SNPs. And almost 50% of the SNPs are validated, which is something very important, means that the SNP has been observed independently multiple times. So you know it's a true SNP.

It could be, many times, since sequencing is not an exact process, if you just look at a set of sequencing traces, you could think that there is a SNP when it's actually just a sequencing error. Now, if the SNP was validated, means it was observed several times by independent investigators, and that gives you the almost total certainty that it's a true SNP.

There are other databases of SNPs. Another very important one is the SNP Consortium Database at Cold Spring Harbor, that offers-- the important thing about TSC is that first of all, all the TSC SNPs are validated. So they basically take SNPs on dbSNPs, then they check them again, to make sure that they're really SNPs. And while doing that, they also look at the frequency of the SNPs.

So what do I mean? A SNP is a polymorphism that substitutes the nucleotide you should have at one location with a different one. So if you look at a population of individuals, you're going to see that the major allele of the SNP, the common one has a certain frequency, so it appears, for example, in 80% of individuals. And the alternative allele appears in 20% of the population.

Now, knowing this frequency is very important, because it allows you, then, to do association studies. For example, to look for a correlation between a disease and this polymorphism. Because if then you observe a second population that is affected by a disease, and you find that in that second population, the alternative allele occurs with a frequency of 40% instead of 20, then that might be an indication that the SNP has something to do with the disease.

But in order to be able to do this, you have to know what is the baseline frequency. What is the original frequency in normal, so to speak, human beings. Yes?

AUDIENCE: I guess the question then becomes, what is the base population is across base population--

ALBERTO RIVA: That's what I was going to say next. Of course, the biggest problem here is that different populations may have different frequencies of SNPs. And this is one of the reasons why SNPs are used for population genetics, because especially in the past, when populations were much more closer there now. If a SNP arises in a population, then it tends to be limited to that population. You're not going to find it in a different population, unless there is some genetic interchange between the two.

So when you look at the frequency of a SNP, it's very important to specify what population you're looking at, because we're going to have an example in two slides. So let's just get back to this in a second. Because I wanted to tell you about other SNP resources quickly.

Haplotype Map project-- this is kind of a new project that is aimed at developing a haplotype map of the human genome. I don't if you had a lecture about haplotypes, about selecting-- you're going to-- OK, so you're going to have it later than this one. But when you hear about haplotypes, just remember that HapMap is a project that is aiming at building a complete haplotype map of the human genome.

And don't have time to go into that now, but it's a very important resource that is really the next step after what TSC is doing, after determining the frequency of SNPs in different populations, that the HapMap project allows you to understand exactly what this can tell you about the evolution of our genome. But this will become clearer in the lecture about haplotypes.

HGbase, another database of SNPs, it's manually curated, so you find-- it's very limited, but you find information that has very high quality. It's all manually verified, and it focuses on the potential consequences of SNPs. So you're going to find a lot of information about known associations between SNPs and diseases.

Alfred at Yale is another very small database, but it has a very high quality, and it focuses on frequency data. And what they do is very, very interesting. They go look at many, many different populations, and especially a population, at small isolated populations from where we are, places like small islands, and Pacific or remote villages in Siberia, and so on. So they actually try to look for isolated populations to maximize the differences in SNP frequency that they're going to find, in order to have a picture of human diversity as complete as possible.

And finally, SNPper that I'm citing because we developed a chip, this is a resource that tries to integrate information from all the places that I sited so far. So it takes information mainly from dbSNP, from Golden Path, from TSC, from Alfred, from HGbase, and it tries to put everything together in a unified view that allows you to look at the gene, find all the SNPs around that gene, see all the features of the SNPs, whether they are different decoding sequence, or the promoter sequence, or whatever. And then look at everything that is known about individual SNPs. And it provides a way of exporting this data in different formats to make it easier to process later.

And I just want to show you one slide from SNPper, but this is a window that describes-- that tells you information about the particular SNP-- this is a [? SNP ?] identifier. And so you can see there is a top part where you have general information where the SNP is, just the position on chromosome 6, where the alleles are, the gene it belongs to, notch 4. And here, it tells you that this gene is in the coding sequence of the gene, and it actually causes an amino acid change at position 319.

It affects protein domains-- this is the list of protein domains that are affected by the SNP. This is the list of some [? matters ?] of the investigators who have observed this SNP-- and it's a long list, so it means that this is definitely a true SNP. And finally under here, I wanted to show you, this data comes from TSC, and it's a frequency information data.

So they sampled 41 individuals from a population of African-Americans, and they found that these are the frequencies for two alleles, 72% A, 20% G. And then they looked at a different population-- these are Caucasian, I think-- and they found very different alleles frequencies. So different that what was the minor allele in the first case is now the major allele.

So this is a very clear demonstration of why it's important to know what population we're talking about when we study the frequency of a SNP. Because if you started-- if you believe these numbers, and then you try to run association studies in the SNP on a different population, you're going to find totally different numbers.

And this doesn't have anything to do with disease-- you're going to just get results that are misleading, because you are not looking at the same population. And the baseline frequency of this SNP and the two populations is very different.

So this is just to show that the advantages of having an integrated view that brings together information from different sources, and allows you to get a clear picture of what the SNP does and everything that is known about it.

AUDIENCE: [INAUDIBLE]

ALBERTO RIVA: Excuse me? This one? Oh, well, it's just telling you that proteins have-- the sequence of a protein is-- well, it contains portions that are active domains, they are the portions of the protein that then physically do something. For example, this domain here is the extracellular domain, is the main that goes outside the cell.

This is a calcium binding domain. So these are structures of the protein sequences that are known to have some function, they are important because they do something. And if you have a SNP that affects one of them, that SNP, in turn, might cause a protein to work-- it can change the function of a SNP-- of a protein.

AUDIENCE: [INAUDIBLE]

ALBERTO RIVA: Well, this is not meant to be an accurate prediction of what the SNP does. And we get so many because all these domains are overlapping. And this information comes from Swiss [INAUDIBLE] database of protein information. And so you see for example, this first domain covers almost all of the protein.

AUDIENCE: [INAUDIBLE]

ALBERTO RIVA: Excuse me?

AUDIENCE: So six would be the maximum number?

ALBERTO RIVA: No, no, it's just that these domains can be overlapping, just because the Swiss people, they annotate the protein sequence saying OK, from here to here, we know that this happens. But-- well, sometimes they-- well, there are some domains that cover the entire protein, or half of the protein, just because, for example, in this case, the extracellular domain, it means that this portion of the protein is extracellular. And then inside that domain, you can have other subdomains like all these that have other characteristics.

So I'm just reporting here a list of all the domains that contain that location, but they can be overlapping. So it doesn't necessarily mean that the SNP affects all of them in some meaningful way. This one is probably the only one that could be affected by the presence of a SNP. Because it's a binding domain, so it might be that it doesn't work anymore as a binding domain.

So don't get confused by this place. Just a list of Swiss [INAUDIBLE] domains that include that location.

Now, the next step I'm going to talk about genes again. And the starting point, when we talk about genes, it's LocusLink. LocusLink is a curated directory of genes from 13 organisms. The word curated here is very important.

So genes are discovered either experimentally, or by programs like GenScan that look at the DNA sequence and tell you where a gene might be.

Then the gene has to be studied in order to know what it does, what all its relationship to other genes and biological processes are. So LocusLink is basically a repository of information about genes, and it collects everything that is known about the genes. So they say their central function is to establish an accurate connection between the defining sequence for locus and other descriptors.

It basically means, you have a stretch of DNA, you know that there is a gene, let's collect everything that is known about that gene. So it gives you information about the sequence, itself, about the functions of the gene, links to other databases, about the gene, different names for the gene, phenotypes that are known to be associated with that gene, homologous to other genes in the same organism, or in different organisms, the location of this gene in several different maps. This is all information that you can find in LocusLink.

And the most important thing-- at least from our point of view-- is that LocusLink provides a nomenclature of genes. No LocusLink assigns a name to each gene, and if you stick to that name, then you're sure that everybody knows what you're talking about. Because this, again, it might seem a trivial problem, but for historical reasons, in many, many cases, genes have lots of different names, even if it's the same gene, people have been calling them with different names, and it's a mess when you try to out which gene is which.

If you stick to the LocusLink nomenclature, then at least you have one way of naming genes, and that's it. So it gives a name, it gives a number, and you can use these as identifiers to look up your gene in other databases, if they use the same nomenclature. And of course, again, it's part of the NCBI cluster, and all NCBI resources use this way of naming genes.

Then, unfortunately, there are other resources that we'll mention later that use a different way of naming genes, and this makes things very difficult when you're trying to build programs to integrate information from different places, because it's very, very hard to know exactly how to reconcile different ways of naming genes.

Again, it might seem a trivial problem, but it's not. And it's also complicated by the fact that, as I was saying before, genes may appear in several different forms-- there are variants of the same gene, there are genes that are very similar to each other. So sometimes they are considered to be the same genes, sometimes they're not. And all these are things that make the naming genes kind of a complex and not a deterministic task.

So UniGene is another resource at NCBI that takes a slightly different approach. It's an attempt at collecting all the GenBank sequences that refer to a region of the genome where a gene is known to be. So essentially, if we know that a certain region of our chromosome contains a gene, then we can go into GenBank and look at all the sequences that fall into that region.

So all the sequences ultimately come from that gene or part of that gene. And UniGene puts them all together in one cluster. And then tries to provide a description of why all these sequences-- of description of the features of all these sequences.

So they're all similar, they all come from the same location from the same region of the genome. But they might represent multiple forms of the same genes, so they're probably not identical to each other. They might come from different tissues, so they might have different properties, and so on.

And again, this is the URL for UniGene. It includes information for 38 organisms. And I think that one year ago, this number was something like 14. So it's growing very fast.

And the interesting thing is, this is an automated process. So LocusLink is a curated directory, means that there are people who spend their days going through gene records and adding information, checking it, correcting it. UniGene is an automated system, so it's actually an automated procedure that looks at all the GenBank sequences and tries to build these clusters based on the location of the sequences.

I've mentioned the fact that it's interesting to study homologies between genes and different organisms. So HomoloGene is a database of all orthologs. So what it does, they take all the sequences in GenBank, they compare each sequence with all the other sequences in GenBank, at least in a set of organisms.

And if they find a good match between the two sequences, then this pair is added to the HomoloGene database. So right now it encompasses 25 organisms. And in these 25 organisms, they have 470,000 ortholog pairs-- so pairs of genes from different organisms that are highly similar to each other.

All these are put into the database. And then if you find there are three organisms that share a similarity relationship, then this, in turn, is marked, because it means that you're finding a match that has an even higher quality. So if you find that organism A shares a gene with organism B, and B shares it with C, If then you find a C shares it with A, then you've built what they call a triplet. And that's a confirmation that actually this gene might be really the same gene that is conserved across all these organisms.

This one is partly curated, partly calculated. So they have an automated procedure that looks at sequence similarity using all the many algorithms to do that. And they give the similarity score. And then they have a subset-- this is not mentioned here. But then, most of these entries in [INAUDIBLE] are also manually curated to make sure that they are really-- that they're really similar genes.

AUDIENCE: [INAUDIBLE]

ALBERTO RIVA: I think it's a Swift-- no, it's part of the [? blast ?] score. Yeah, they have a threshold of something-- I don't remember. But they give it the score in addition to all the other information.

OK, Ensemble-- this is not part of NCBI. This is something that comes from Europe, from the EMBL-- European Bioinformatics Institute, the Sanger Institute. It's something that is pretty similar to LocusLink in scope. Again, it's a software system for the automated annotation of genomes-- it's basically means it's a system that discovers genes and tries to find as much information as possible about these genes.

And then all the information is available through a search interface. It's limited to 10 organisms, but it provides a lot of information about the genes in this-- about this organism. So it provides information about genes, about proteins, diseases, SNPs, cross-species analysis, microarray data. So it's essentially a combination of LocusLink, dbSNP, HomoloGene, and a few other things.

It has a very powerful data access interface. It's actually very, very nice, very easy to use. So you can do queries on this huge database in a relatively simple way.

One of the biggest problems with this system, at least from our point of view, is that it uses its own way of naming genes. This is essentially what I was referring to before when I was saying that not everybody uses the LocusLink way of naming genes. They have their own alternative scheme for naming genes. And going from one to the other is sometimes tricky.

There are links between the two databases. But of course it's not-- they don't necessarily match very well. What else?

OK, and finally, a few words about gene regulation. So gene regulation-- of course, it's almost needless to say, it's an extremely complex mechanism. Our understanding of how gene regulation works is still very limited.

When you hear about microarrays, about the concept of gene expression as measured by microarrays, gene expression is the most visible consequence of everything that is in this complex mechanism. So what you see is that under certain conditions, a certain set of genes is highly regulated, is highly expressed, or under expressed, and so on.

But this is a consequence of the fact that there is a very complex machinery behind it that determines which genes are active or not, and how much, in different conditions. And this is actually a system that integrates a lot of different factors that might include the following, in no particular order--

The tissue, we know very well that the set of genes that are expressed in one tissue is very different from the set of genes that are expressed in another tissue. Developmental stage, genes that are expressed during the development of the embryo, for example, are not the same that are expressed in an adult organism.

The time-- the time can mean either a time of day, for the case like the circadian rhythm, there are genes that are expressed in the morning and not in the evening. Or time at a larger scale, there are processes that take years to complete, like puberty, for example.

So there is this regulation mechanism is able to work at very different temporal resolutions. External signals, of course, all response to external stimuli. And it also depends on the expression state of any number of other genes, because genes regulate each other through feedback loops and so on.

So again, it's a very complex system. We're slowly working to try to understand how it works. So what we have for now is some understanding of what transcription factors-- the transcription factors are proteins that bind to the upstream regions of the genes, and are able to control their expression, their activity. Because transcription factor, because usually the most common case-- these bind to the promoter region of a gene, they combine with each other, forming complexes.

And these complexes then activate the transcription machinery, that then gives rise to what starts everything else. And in the end, you get the gene is expressed, because it was latent in the protein is produced.

And the transcription factors, as I was saying, don't act alone. They have to interact with the target gene, but they also interact with each other in a combinatorial fashion. What this means is that, looking at the individual transcription factor is usually not sufficient to understand what it's going to do. Because the same transcription factor in different combinations with other transcription factors might have different roles.

So what we need to look at is the pattern of transcription factors that binds to a certain gene. And that, in turn, will determine the spatial, temporal, dependent expression of the target gene.

And again, we are still doing the very early steps in the process of trying to understand how these patterns are actually structured, how they work. So what we-- OK, sorry.

So the first step, again, we're moving the first steps. The first thing you need to do is need to be able to reliably identify which transcription factors bind to a given gene, and where, exactly, in the promoter region of the gene they bind. And transcription factors bind to locations that are called transcription factor binding sites. They're small stretches of DNA that are recognized by the factor.

And so if you know where the binding sites are, you have a first idea of what factors bind to this gene, and how they may be arranged spatially. So if you know that two factors have to interact with each other, probably their binding sites will have to be close to each other. Or, at least, let's say if you find two binding sites that are close to each other, there is a very high chance that two factors will interact.

And it might be that when they interact, they act in a certain way. Where they don't interact because they're far apart, and they act in a different way. And so knowing the map of binding sites, and the promoter of a gene is something that can give you some initial information that it can build on.

It's still something that's very hard to do computationally. The ways that [? truth ?] that people have been using to do this are usually based on pattern matching. So the binding site, as I said, is a small stretch of DNA, usually goes from five to about 20 or 25 base pairs. So they're really short.

And they're characterized by [? concentric ?] sequences-- in general, they are not very [? concern-- ?] not very precise. So it's essentially impossible to look at a piece of DNA and say, well, OK, I'm sure that this location here is a binding site. So you can try using deterministic methods, just looking for instances of the motifs.

Sorry, thought I had something on this. But you can look for instances of the motifs using either deterministic methods, or probabilistic methods, pattern matching, there is lots of things that you can try. And in almost all cases, people rely on TransFIC-- TransFIC is the largest available database about transcription factors.

It's a database that provides information on the factors, themselves. It provides examples of their binding sites. And it provides descriptions of their interactions with genes. And the important thing is that most of the information in TransFIC is experimentally validated.

So for example, the binding sites, these are binding sites that have been observed experimentally. So you can actually trust the fact that particular piece of sequence they give you is the binding site for the transcription factor in question.

And so in the end, without going into too much detail, what you can do is can take these binding sites, you can use them to train your favorite pattern matching method, and then you can try scanning new sequences looking for binding sites. And this is one of the things that we're currently working on at CHB-- there are various ways of doing this. And again, it's a rather difficult problem, from a computational point of view, because these patterns that you have to look for are not very specific. They are not very clear.

On the other hand, doing it computationally-- sorry, doing it experimentally is very slow, very expensive. So you can only do that for a small number of genes, and a small number of factors. If you have a method like this, if you have a computational meter to detect binding sites that works well, then you can think about doing this on a large scale, looking for example, all the binding sites for a certain factor in all the human genes. And that will give you a very interesting picture of everything that might be regulated by that factor.

So we're not there yet. This is one of the things we are working on in our lab. And it's going to take a lot of work but the rewards are potentially very interesting because this is something that will then allow you, if it works, to automatically build the networks that describe how genes regulate each other. And that is something that, of course, has a lot of potential interest.

OK, we've talked about gene expression, we talked about microarrays. You might have already heard about these things, but I was just going to list the main sources of available microarray data, public microarray data. So for example, again at NCBI, GEO is a database, Gene Expression Omnibus, is a database of gene expression and hybridization array data.

It offers 12,000 experiments, essentially, 12,000 hybridization experiments on over 500 . Platforms so if you're interested in doing some form of data analysis on microarray data, and you don't have the time or the money to-- sorry-- to do your own microarrays, you can go to GEO, and you have 12,000 of them to choose from. And they also offer a very powerful interface to search-- since microarray data sets are very large, they include thousands of measurements.

They provide a very useful search interface that allows you to select the data sets we're interested in, and to extract data from these data sets and look at, for example, the behavior of the same gene in different experiments, or different genes in the same experiment. And there are lots of different queries that are common when you work with microarray data.

The Stanford microarray database, again, is a repository of all the-- of a large number of micro experiments performed at Stanford, and a portion of these are public.

NCI60, again, from Stanford, is a famous data set that includes gene expression profiles for 60 human cancer cell lines. And the information on drug activity correlated with gene expression patterns. So they measure how the gene expression patterns change when these cell lines are subject to different drugs.

Other resources for gene expression are found in different PGA projects, PGA are programs for genomic applications, they are are large projects managed by the NIH. So the [? tracks ?] PGA, for example, offers 565 microarrays from mouse and rat models of sleep, infection, hypertension, pulmonary disease. The Hopkins PGA, again more than 500 microarrays from several human diseases.

Cardio genomics provide microarray data on mouse models of cardiac development and signal transduction. And finally, human gene expression index-- these are just some of the most important most useful public resources of microarray data.

OK, I'm going to go through this final part very quickly because I'm almost out of time. And if you'd rather stop me with questions, or if there's anything you would like to discuss about what I said so far, we could stop here, or I could just run through this last portion quickly.

So this last part was about the last step in the process from proteins to phenotypes. I was going to talk about protein databases. The situation in protein databases is a bit different from what we've seen so far. The protein world is much more complex than the DNA and RNA world for the reasons that I've explained at the beginning.

Some of the reasons are that proteins interact with each other in very complex ways. They combine in three dimensions, they catalyze chemical reactions. They have a behavioral that is much harder to describe in [INAUDIBLE] terms than everything else we've seen so far.

So what protein databases give you is usually information about the sequence of a protein, and that's the easy part. The known or computed three dimensional structure, the known or inferred function domains. And ideally, also, the functional protein, what the protein does in different conditions. But again, this is-- we're getting to the area where things start becoming hard to formalize and to represent in a computational system.

So as a consequence, protein databases, first of all, tend to be older, because they were started earlier than genomic databases. They are less integrated, they are less complete. Nomenclature is much less standardized. So it's harder to work on protein databases than with all the other resources we've seen so far.

The biggest database is SwissProt, 120,000 sequenced centuries, 9,000 human proteins in SwissProt, which is a pretty small number if you think that we already have complete information in Golden Path about 20,000 genes, and each gene is known to, on the average, code for probably two or more proteins. So these are the proteins for which we know something, and they're very, very, very small number compared to the total number of proteins that are thought to be in our cells.

It's composed of a core set of data elements, this sequence, the references, taxonomic data for this protein. And then our [INAUDIBLE] about the functions of this protein, domains and sites, the structure, similarities, association with diseases, variant forms of the protein. And again, it's hard to link this database with LocusLink or UniGene, but is its own identifiers for proteins.

But don't need to go into these problems now. This is a graph that shows you the growth of SwissProt in recent years. And as you can see it's growing but at a much smaller rate than GenBank or other resources like that.

We have databases about the three dimensional structural proteins like PDB, different visualization options. MMDB is essentially the same thing, but implemented at NCBI. PFM at the Sanger Institute is a database of protein domains and protein families. They look for domains in the proteins, and then they look for similarities between proteins on the basis of the domains that were identified. They use similarity measures, they use hidden Markov models.

Again, they have a curated portion with a small number of protein families. With the notation there is a high quality. And then there's a second portion of PFM that has smaller families of lower quality.

This is an example of a display, a PFM display, of a protein with all the different domains that were found in the protein with the tails here. So give it this nice graphical display.

I'm going to skip protein interaction databases. And I want to get to the end. We're getting to the phenotype and to the spectrum, finally. And there's just a couple of resources that have to be cited because they're extremely important. One of them is OMIM-- OMIM is a catalog of human genes and genetic disorders. Again, hosted by the NCBI.

It's basically a collection of text articles that talk either about a gene or about a disorder, and they're linked with each other. So if you're looking at the entry for a gene, you can find a description, mechanical feature, the function mapping, and then you can find all known correlations between that gene and diseases, allelic variants, so all known polymorphisms of that gene with the corresponding clinical outcome if there is any. And then you can also go the other way around.

It has 14,000 entries. Again, these numbers are probably larger by now, because again this is a graph that shows you how-- it is not very up to date, but you can imagine that it's been growing at least at this speed or faster since '98.

And finally, PubMed, you probably all know what PubMed is, a database of citations from the biomedical literature. It contains 12 million entries starting from the mid-'60s, and it provides references, abstracts, linked to online resources. Full text articles, in some cases, supplementary materials, and it's one of the most used resources in this field. They claim they receive 30 million searches per month.

OK, one last thing-- gene ontology. Gene ontology is something that stays at a slightly higher level, above everything that we've seen so far. The idea of gene ontology is to build a dynamic controlled vocabulary that can be used to describe biological concepts. If you look at something like OMIM or PubMed, you're going to find a textual description, for example, of a disease that references concepts that need to be precisely defined so that we all know what we're talking about the same thing when you use the same word.

And the purpose of gene ontology is to try to do this in at least three domains-- molecular function, biological process, cellular component. So it's organized in three taxonomies, and each taxonomy contains concepts and sub-concepts and so on, that try to describe everything that is known about molecular functions, molecular biological process, and several components using a standardized nomenclature. So that when you want to refer, for example, for to a certain component of a cell, instead of just saying its name, you can cite the gene ontology term that describes that component, and everybody else will be able to go to gene ontology and see what's the exact definition of the word you're using.

It's a work in progress-- still very far from being complete. It has all the usual problems that occur when you're trying to build taxonomies that it's very hard to formalize things that come from natural language. So it could find exact definition of all the terms that people use, especially in this field is very hard. But this is where they are now, and it's a work in progress, so it will keep growing in the future.

And this is a view of taxonomy, for example, for biological process. If you are talking about site communication, then response to external stimulus is a subclass of communication. Response the biology stimulus, the first response is a subclass again of all this. And if you want to talk about the immune response, you can cite this biology term, and everybody will be able to go to gene ontology and see exactly where this term is in the taxonomy of concepts about biological processes.

OK, I think we're out of time. Well, just a conclusion slide that I'll just let you read, because I think it's just repeating what we're saying so far that we are drowning in data and converting this data into knowledge is not easy. We need automated tools to access this data, to make sense of it, to convert into formats that we can use.

And of course, this is a challenging task, because as we saw, biomedical data covers the whole spectrum of knowledge representation and management techniques that we know about.