The following content is provided by MIT OpenCourseWare under a Creative Commons license. Additional information about our license and MIT OpenCourseWare, in general, is available at ocw.mit.edu.

GEORGE CHURCH: OK, welcome back. I'm sure you're all dying to know the answer, so y'all came back faster than usual.

[LAUGHTER]

So how do we get from the first top motif score to the second one? And I'll show you, just to spice this up, I'll see two different algorithms they actually, historically, done and the 100-fold improvement in speed and also in accuracy that comes from the change.

So the first way, and you already can see negatively predisposed to this, but if you look at the motif, the winner, if you will, the first time around, it has certain base positions which are particularly high-information concept. That is to say they really dominate, and they probably are critical to finding a motif. If you had a way of, say, knocking out one of those from the sequence, from all the sequences which contributed to this motif, then it would greatly reduce your chances that you would find it again.

And so that's what we are going to do in slide 47 is we're just going to go through and pick on one of those bases and turn it into an X. So an X doesn't really match any of the weight matrices. And so whenever you have an motif that overlaps it, it won't have a good score. And so you won't build up-- you won't have this transition. The Gibbs sampler won't go in that direction.

Now, this has a couple of disadvantages from a accuracy standpoint in that there may be some motifs that you really like that overlap the original motif slightly. And you'll miss those during the sampling. So an alternative way of looking for these sort of things, rather than taking the best continuous sampling in this Xed out version, instead, what you can do is maintain a list of all the motifs you've found up to this point. In this case, we have just one.

And now, because we were using AlignACE to do this multi-sequence alignment by sampling-- but now, as you're going along, building up, initially, a random motif, you compare it to the first winner. And you say, is this random motif, or this motif that's emerging out of this random process, does it look at all like the previous one? If it starts to look convincingly like the winner before, you know where it's going to go. It's just going to get more and more like that. So you might as well quit early.

And so that's what you do is now you haven't Xed out any particular base. All the information is there. You can take any kind of motif that expands and changes the columns and so forth. And if you're building up towards a motif you've seen before, you can reject it.

This process has the dual advantage of now allowing you to get overlapping motifs that might have a different enough column structure or differ enough weight matrix or are slightly offset so that it really is a different motif. You can find those. The algorithm that you're using to compare these we'll use a couple of times today. We'll call it CompareACE for comparing these consensus elements, or these weight matrices.

And so it not only improves your ability to discriminate related but statistically separable consensus elements but it also has about 100-fold increase in speed. Because you can stop 100 times earlier in this motif sampling, which goes-- once you lock into a motif, you go and you go until you really get the best possible score. But now you can reject these weaker ones earlier on.

Now, you may have been wondering all along, you probably have an intuitive feeling for what the MAP score is-- and I'm going to take-- and at the end of this, you're not going to be able to rederive this from first principles because I'm not going to go into it in that depth. But I want to expose you to some of the terms that are involved in this maximum a posteriori score.

Of course, the hero of any kind of scoring function is the weight matrix, the actual number of counts of As, Cs, Gs, and Ts you have at every position in every column in this matrix. Now, remember, we've critiqued this already, that the typical weight matrices, there is no codependence of columns the way there was in RNA secondary structural or in CpG islands or other Markov chains. These are independent columns.

So the key player, the hero here, is f sub jb. This is the weight matrix. And this is not a frequency, but this is the actual number of occurrences of each base-- A, C, G, or T, which is B, at position j in the column matrix. These can be the active columns, j going-- and then the number of occurrences is just the sum of those occurrences.

We've already been talking about how the width of the motif can include some columns that are basically on and off, columns that you believe are significant and those that aren't. So the number of columns, c, is less than or equal to the width. So for example, when we were doing the GCN4 example, we had a width of 10 or might at one point expand to the width of 11 with 10 active columns, c equal 10.

You'll recognize-- here's the star, the f sub jb here. And you're adding these pseudocounts here, these betas, which you remember that every time that you have a danger that you might, because you have a limited database that you're looking through, a limited number of actually observed sites, you could get a number of sites equal to 0.

And you don't want to have zeros in there. Because you're basically acknowledging, because I did lambda sampling, it could have been 1. If I had sampled one more, it could have been 1.

And so an estimate might be that you add another pseudocount. And this can be represented here. You don't want to have-- and these gammas you can think of, they're kind of like factorials. And you're taking products. The pi is products.

And then I may have mentioned that you might want to take into account the background levels of the bases. If you get, say, a motif that's just a string of As and you're doing it in a genome that's very AT-rich, then you want to account. It's less surprising if you find a string of As in a genome that's GC-rich. And that's what this g sub b is the background genome frequency for base b.

Now, in a double-stranded DNA, the frequency of As is going to be equal to frequency of Ts. But a single-stranded RNA virus, for example, there's going to really be a very independent set of backgrounds for Gs, As, Ts, and Cs, or Us. So this gives you some flavor of what's in the MAP score.

A greatly oversimplified version of this detailed slide 49 is in slide 50-- laughably oversimplified-- is it's basically the overrepresentation of these sites is what we're talking about. It means that these sites, you're giving a higher score if there's an overrepresentation in that learning set.

It tells you nothing about the rest of the genome. It could be they're overrepresented in the rest of the genome, too. And that's what we're going to go into next. But you get a bonus for the number of aligned sites and the overrepresentation of those sites. That's what the MAP score is.

STUDENT: Does the background [INAUDIBLE]?

GEORGE CHURCH: Hmm?

STUDENT: Does the background [INAUDIBLE]?

GEORGE CHURCH: The background is ignored in this oversimplification, but it's explicit in slide 49. You really should ignore this and think more about the previous one. But the main point that I'm making is the overrepresentation is just half of the story. The other half is the specificity.

In other words, if it's present in your learning set or your enriched set, you want to next ask, is it present in the rest of the genome? Because if it's present-- and maybe that's what you mean by background-- if it's present in the rest of the genome, then that's not great. And so what we're going to do is this is an example of running through-- after you get the first motif-- running through lots of other motifs.

So the best motif of all for a larger set than the ones we were looking at. We were looking at seven, but there's 116 of these in [INAUDIBLE]. When you run through the whole thing, the top MAP score, the one that's most overrepresented, is this A-rich one.

Now you want to ask, is that specific to the amino acid biosynthetic genes? Or is it found all over the place? And we're going to get to that, how you measure that, in just a moment.

The one we were kind of highlighting here is in GCN4 is kind of modestly-- this is not a rank order list. This is kind of in random order. But we'll show how you can do the rank ordering, as well, how you can order this.

But you see all kinds of motifs, some that are stretched out, different compositions. So to evaluate motif significance, we have these five examples that we'll go through. There's the specificity that I've been talking about and will be the subject of the next slide. That little arrow means the slide that's due here.

Group specificity-- is this specific to the group that you found by clustering, or is it all over the place? Functional enrichment-- we talked about this a couple of times. Are the genes that you're finding in the cluster or the genes that you're finding this motif in front of, are they enriched by some fairly objective criteria?

Is the motifs you're finding in a particular position in the upstream elements? Because they have a position in the promoters or enhancers. Does the motif that you found have interesting symmetry properties, as you might expect from proteins, which bind as multimers? They might have inverted or tandem repeats, where the elements either point towards each other or in tandem.

Is the motifs that you're finding, are they related in any way to motifs that were known before by more complicated biochemical and genetic assays? So the first one, the group specificity, in order to ask whether the motif you found in the small subset of the genome is present in the rest of the genome, we need a way of scanning.

Now, when we introduce weight matrices in the multi-sequence alignment lecture and we say we would put off the motif Gibbs sampling until today, we already introduced one really trivial way of scanning the genome, where you basically take the weight matrix, move it in each position, and you do a simple sum. That's basically what this is is a simple sum.

But we're taking log ratios of these counts. Again, the hero, the counts, just like the f-- so bj before, now it's n sub l b, slightly different nomenclature taken from different articles, but it's the same idea. This is a number of occurrences of base b at position l. This is the weight matrix as counts, not as frequencies.

And in the denominator is the number of occurrences of the most common base. Now, this could be b, or it could be some other one. But this is the most common one. This is going to tend to be larger than or equal to the ends of lb, which is the weight matrix.

And you're going to sum over l, over the length of the binding site. This was w in the previous one. And you're going to just scan this along the entire genome, stepping it over one base at a time and coming back on the opposite strand. And that's going to be ScanACE.

So you've got AlignACE compares to ScanACE. Now we're going to scan it to see for specificity. Again, you've got these 0.5's. You can think of these as pseudocounts. Keep the zeros out. You don't want to have a logarithm of 0.

Now, let's walk this through a particular biological data set. This is a cell cycle data set going through two cycles, cell division cycles. here. There are points, 15 significant time points, along the horizontal axis.

And this is a particular cluster. Out of 30 different clusters, this particular one has a peak just before the S phase, the phase and the cell cycle where you're trying to get the replication of the cell at the S phase is where you actually synthesize a new set of DNA molecules. You duplicate it, just as we talked about in the first lecture.

And actually, since you've recorded a time series through two cell division cycles, you expect there to be periodicity. Or genes that are acquired in the first S phase, you expect them to be acquired in the second S phase. That's the underlying thought behind designing this experiment is you would synchronize all the cells.

Normally, cells are all over the place. Some of them are in S. Some of them are in M, which is where the metaphase chromosomes separate. But in this, you synchronize them all up by a method we'll talk about in just a moment.

And then this is that diagram where we have the number of standard deviations from the mean. This is a normalized signal of the RNA expression. On the vertical axis and horizontal axis is this time series categorically described as G1-- gap 1-- synthesis, gap 2-- metaphase-- mitosis, and so on.

Now, what do we learn from this particular cluster? This cluster has 186 genes in it. That means the RNAs for those 186 genes were in a nice envelope. It doesn't mean that they're strictly couldn't be 185, 187. There may be some outliers on the edges. But that's the number that we're going to be doing these calculations for.

The ways we're going to evaluate it is, first, whether the functional categories make sense. Is there an enrichment for a particular functional category? You may have already, those biologists among you, may have already had a hypothesis of what functional categories should be enriched. If it's going to peak, if these are the RNAs that are going to peak just before S phase, just before you need them for DNA synthesis, maybe these can encode genes that are involved-- genes [? possibly ?] involved-- in DNA synthesis.

Sure enough, that's the most striking observation is that in this database, this MIPS database of functional categories, you have 82 genes that are described as involved in DNA synthesis. And of this cluster that are co-expressed at peak, at S, you have an overlap of 23 with that. And that may not sound like a huge overlap-- a few-- but it's very statistically significant. It's 10 to the -16 is the probability of that occurring at random, out of the 6,000 or so [INAUDIBLE] genes, that having this overlap of 23 is very significant.

So that's the first. And we'll, in just a moment, show how we did that calculation. But that's your first test. There is a functional category enrichment. Next, you find the motifs. You use the AlignACE. You go through. You find the top motif. It's MCB. You go and you find the very close second-highest motif. It's SCB.

These are not chosen by hand. This is all done algorithmically. The only input-- there's no literature input except for checking these functional categories-- for finding the motifs, it's just the microarray data and the sequence upstream from the genes that come out of this cluster. That's how these were found.

Now, they have to have names. MCB and SCB, we could have just called them x and y. But these names do mean that the CompareACE score to something that was in the literature is good. But a very profound accomplishment of this is that now-- unlike the literature, where it's rather challenging to find the connection from a conclusion, like this motif is likely to regulate this gene is likely to be enriched in this class of genes-- here, it's directly traceable.

You can see the logic that connects this motif MCB to this cluster via Gibbs algorithm. And this cluster is traceable back to the RNA profiles on the microarray. This is all a very simple, comprehensive study.

But now you want to ask, is this motif-- we know that it's got a high MAP score. It's highly enriched. And it's highly unlikely that a motif this strong would have occurred in this size cluster of genes. But what you want to ask, is it specific? This is the thing we've been putting off for a little while. Is it specific?

And if you look at the 30 clusters when you cluster this whole set of genes that vary during the cell cycle into 30 envelopes, this particular one, envelope number 2, cluster number 2, which is displayed in the upper left, down in the lower left, you can see that all the MCB motifs, almost all of them, are found in cluster 2 when you use ScanACE, and very few of them are found in the rest of the genome-- similarly for the second-most impressive motif by AlignACE MAP scores, SCB. And it's also specific to cluster 2.

The fact that you're seeing this non-random enrichment for functional category, this non-random enrichment for a motif, this non-random specificity of that motif and a second motif, kind of tells you that everything is working, that your RNA data collection is working-- which, you may spend a lot of money to get to this point, you should be gratified-- and the clustering is working, and the motif finding and specificity scores, all this is working. It doesn't mean that it's absolutely perfectly tuned and everything, but it's giving you feedback that you're taking a step in the right direction.

Similarly, the position of this motif in the promoter is non-random. You see this little spike that's coming up just before the-- it could be the transcription or translation start. In this case, the ATG is the translation start. And it's non-random.

How do we measure each of these things? How do we get this? Well, before we get that, I'll give you two more examples, same format, just to show you that you get different motifs when you go to different clusters-- two more clusters.

The next one is also periodic. It has a peak now slightly shifted to the right from the previous periodic function. And it repeats at exactly the same periodicity as if they're part of the same periodic function, which is exactly how the experiment was designed.

The difference between this and the previous one is now, the two top motifs are not previously known motifs. It doesn't mean they're any worse. But they're new ones. And the way you evaluate whether they're specific is the same way we got the specificity for. Now they're in cluster 14, which is this cluster up in the upper left.

And both of them are about as specific as the ones in the previous slide. The functional category is not as impressive. It's 10 to the minus 6 instead of-- sorry, 10 to the minus 4, no 10 to the minus 6, the previous one's 10 to the minus 16. Now, this is still statistically significant. But it could mean that this particular way of functional categorization which the curators use may not be ideal for this particular regulatory mechanism, regulatory regulon.

So this may be a discovery both of a new regulatory set and of two new motifs. But in order to establish that, you'd need some experiments to really, say, knock out these motifs and see what the consequences are.

Now, the third cluster illustrates yet another set of ideas. Here, even though the experiment was designed specifically to enrich for the most abundant-- or sorry, for the most periodic gene expression, there were inevitably features of the experimental design which were not periodic. In particular, when you synchronize the cells, you force them all to be in synchrony for the cell division cycle, you did that by taking a temperature-sensitive mutant in the cell division cycle, say CDC15 or 28.

And that temperature-sensitive mutant, that requires that you raise the temperature to shut down the function of that gene by unfolding the protein. And so you have a temperature shift to allow them to go back into the cell cycle. So you're going from high temperature to low. That's one thing. And so that temperature essentially decays rapidly. And then you have the residual of that going out in time.

In addition, there's all the physiological effects. You had all these things kind of waiting in this funny physiological state. And then that decays with time. So that's not cyclic. That perturbation is a linear or decay.

And sure enough, you find examples of clusters which are not periodic. This one peaks in the second cell cycle but not in the corresponding point of the first cell cycle. And in fact, most of the 30 clusters, when you divide this up into 30-- the entire expression space up into 30-- are like this. They're not periodic.

But that's OK. Because what you're looking for is clustering, as if these were different conditions or different time points. It doesn't matter what it is. Because they are coexpressed, so going up and down together, possibly during serendipitous factors. But you can still apply the same criteria for asking whether you're impressed with this cluster or not. Does it have enrichment for a functional category in the upper left-hand part of 556?

And wow, it really does. This is the most impressive one of all 30 clusters. It has a probability of 10 to the minus 54 that you would find this degree of overlap between the functional category-- think of this as a Venn diagram of overlapping circles-- the overlap between the class of ribosomal proteins and the class of this particular cluster, which is not periodic, is amazingly significant.

In addition, you find two motifs. The top two motifs are highly enriched. That's what the Snyder information content logo means-- and is highly specific. That's what the bottom line means. This is-- it's present in cluster 1 and very little in any of the other clusters by ScanACE using the motif matrix.

So these are three clusters, each with a different story. The first one was two known motifs. The second was two unknown motifs and possibly a new functional category. The third one is a whopping match to a functional category, one known, one unknown motif, and the whole thing non-periodic, even though the experimental design was periodic.

So now we've shown that you can quantitate all these things that are often casually treated in the discussion section of biological papers. Here, they've all been treated quantitatively. But how do we do that? What is the algorithm behind each of these things?

We won't talk about how we measure periodicity. But you can imagine you can measure the periodicity. And we have. You could ask specificity. How did we measure the specificity and the functional assignments? It turns out that's almost the same statistical function we use for those two things-- functional assignments, group specificity.

Positional bias is a different one. And CompareACE we can use not only for looking for previous motifs, as we did in the AlignACE algorithm itself and as we do as we want to look through databases of motifs, we can also look as the motif looks symmetry on itself. So this is how we do each of these.

We have a choice. When we ask whether the intersection of two subsets of all of the possible genes-- let's say our cluster and a functional category or a cluster and all the best hits with ScanACE-- if those overlap in a significant way, we can think of that as sampling from a population.

The question is, are we sampling with replacement or without replacement? It's an easy thing to get confused about. And I urge you to just look back at the definitions of these offline. But there actually-- a mistake was made in the literature by an author who should have known better because he got it right the first time and wrong the second time.

But the correct use-- and in fact, in widespread use-- is the hypergeometric. Because we are actually, here, sampling without replacement. When you do that, the two sets, the two subsets of the big set-- the big set is n, and the two subsets are s1 and s2-- you have these combinatoric, this simple combinatoric, where you have s1 choose x where x is the intersection between the two sets. And

This will be much clearer in the next slide, where we have a kind of a diagram to go with it. But this is the chance of getting exactly x. In the next slide, we're going to show how we need to consider the possibility that it could be x or larger.

Now, so this is the diagram. n is the total number of genes in [? yeast, ?] somewhere upwards of 6,000. And then the subset 1 might be the number of genes in the cluster that you got out of your microarray experiment. And s2 is the number of genes found in the functional category. This is the MIPS database.

How surprised are we that we found x as the intersection between those two sets? Well, let's say x were 1, and the two sets were about 100 each. That's not too surprising, right? But that hypergeometric formula, if you plug in 1, it's very surprising that you got exactly one.

But the reason is because it could have been bigger than-- we're saying that it's significant that they overlap that much. Well, if it's 1, we have to consider 1 or greater. Because we're basically saying it could be 1 or greater. And so what you have to do is do a sum from 1 up. And what you'll find is that that's very likely, not surprising.

On the other hand, if we had a very significant overlap with these two, then the hypothesis that these two are very related-- in other words, that you've got an enrichment, that your cluster is enriched for this functional category because you've got, say, 100 in s1 and 100 in S2 and the overlap is 99, then you would have been surprised by 99, and you would have been surprised by 100.

But both 99 and 100 together are still very rare. And so you're surprised. And so the sum has to go from whatever you've got on up. And that's what this is. And that's easy to forget, too. People might just say, oh, this particular intersection is surprising. So you have to have that sum.

Now, I'm going to go from slide 59 to 60. And there's going to be relatively, graphically, very little difference. But it's a radically different thing that we're doing. Now we're doing group specificity score. This is the motif you found in the cluster s1. You looked through s1. AlignACE found your motif. Now you want to ask whether that's specific or not.

So you search through the entire genome, and you pick top 100 matches. And those are upstream of the genes in s2, subset 2. If there's a huge overlap of s1 and s2 we'll call x, then you're going to be surprised. And so again, you take the sum over this hypergeometric distribution. And if that's a small number, a small probability, then that's a measure of how surprised you are. So if that's 10 to the minus 6, then you're very surprised.

Now, those were hypergeometric. But positional bias now, it is binomial. And you should be able to remember the binomial is this combinatoric term where t is the total number of sites, and i is the amount you're surprised. So m is the number of sites that are in the most enriched window.

Now, you can take a window any size you want. If you make it too small, you're going to get sampling statistics. If you make it too big, it'll include the entire 600 base pair non-coding region. So you can try a bunch of-- you can try different windows. But basically, what you're looking for is how surprised you are that you have m or more sites in that window.

If you're surprised by 10 sites, then you'd be even more surprised by more than 10. So you have to take the sum. So it's a sum just like the previous hypergeometric ones. But now it's over a binomial. And remember, the binomial is this combinatorial term, a probability to the i power and 1 minus that probability to the total minus i power. So this should be very recognizable. This is the chance that you have enrichment in a particular part of the promoter.

Now, if we can compare motifs-- we've mentioned this already-- we use it in the AlignACE algorithm itself to cut our losses when we start refinding the same motif again. We use it to find whether there were similar motifs. And through experience and training sets-- and you can find that, just kind of like a correlation coefficient, the CompareACE score, as it gets closer to 1, is more and more believable. And about 0.7 is where you get statistically significant matches to other motifs.

And here's an example where you can actually treat the distances to other motifs, so similar to other motifs, where 1 is perfectly similar. As you get along the diagonal, any motif is similar to itself, by definition. And you can build up a little matrix of similarities of motifs. And then you can do hierarchical clustering of motifs.

And you can, if a and b are sufficiently close together, then you might consider that they're the same motif or they're transcription factors where the transcription factor that binds to that DNA motif may be related to the protein sequence level. These are predictions that you might make from this kind of clustering based on comparing weight matrices.

Now, if you compare a motif to itself, what does that mean? If you just compare it to itself over its entirety in the same orientation-- that's what the previous one-- you'll get a comparative score of 1. However, if you flip it-- remember, DNA being double-stranded, unlike proteins-- when you flip it and compare the weight matrices, now you're asking whether it has twofold symmetry.

And this is another very profound connection, I think, between the weight matrices, which are kind of a summary of an alignment of many sequences of evolutionary significance or, in this case, regulatory significance. But it's a weight matrix of aligned sequence. That is actually directly related, conceptually related, to a very different thought, which is that the three-dimensional structure of the protein-nucleic acid interaction has some symmetry in it.

If you have a protein dimer or a protein domain that's duplicated, if you have a cell symmetry to a reverse complement of a motif, that means, in three-dimensional structure, the two protein motifs are related by a dyad symmetry. That means a twofold axis where you rotate 180 degrees in three dimensions.

On the other hand, if the element, if the halves of the element or thirds of the element are related by a direct translation in the motif space, in the multisequence alignment, then that means that you have a direct repeat of DNA-protein interactions where the helical translation and rotation of the axis is reflected in the protein DNA. So anyway, there's a connection between motif matrices and three-dimensional structures.

And here's how it plays out when you do a CompareACE where you actually go and you compare, column by column, the weight matrices of motif 1 with itself in reverse complement. And you can see these three PRRs are in [INAUDIBLE] taken from bacterial genomes are very significant when you compare it to its reverse complement. That means that, very likely, there's a protein dimer or maybe a closely sequenced related heterodimer which binds 180 degrees symmetry.

On the other hand, here, when you compare CPXR to itself reverse complemented, a very poor AlignACE score. It means it doesn't have this dyad symmetry. However, if you took the two halves and compared them-- I don't show it-- but no doubt you would get a very strong AlignACE-- a CompareACE score between the two halves, indicating a direct repeat in sequence space and sort of a helical repeat in three-dimensional structure space. I think this is a very powerful connection between these two. And that, of course, can be quantitated.

Now we want to say, behind the scenes, all along, you've had to have some confidence in what the AlignACE scores meant. And you do this by doing a test set. A test that has to be composed of negative controls and positive controls and very large set of functional categories from which we've shown a few examples in the context of negative controls, positive controls.

So negative controls can be randomly selected genes. And you want to try different cluster sizes to see the effect of cluster size on the whole algorithm. You might be able to predict this completely theoretically. But it's very gratifying, whether you can or can't, to run it through exactly the same algorithm, the same software, with randomly selected sets. Now, this is very expensive. Because you can generate-- you need to generate-- a lot more randomly selected sets than the actual test sets.

And then for positive controls, there are actually relatively few of these. These are cases where you have really well-defined transcription factors, which have to have the additional fact that they have five or more known sites. Because you need to have five or more sites in order for AlignACE to get a grip on the problem and produce a nice multisequence alignment.

OK, so let's go through, first, the results of the functional categories-- 248 functional categories-- from these different databases and then go through the negative and positive controls. So here are some of the friends that we happen to find-- now, this is all done from functional categories. This was not done from microarrays. But here are some of the friends that we found in this cell cycle microarray data.

RAF1 was the ribosomal one. GCN4 we've seen before. And MCB was the one that was in the S phase. And you can see these have been ranked. And remember, so you could rank them by three different methods-- by the MAP score, which is the unlikeliness of finding this good an information content motif in the learning set. It doesn't tell you about specificity. That's the next column to the right.

There's MAP, the specificity score, which means that it's present in that functional category and not in lots of other parts of the genome, and then position the bot. And remember, that was done by the intersecting Venn diagram hypergeometric.

And then the positional bias-- that was the binomial-- is how non-randomly positioned is it in the promoters? And so this is ranked by the specificity. And you can see RAF1 is very specific to that particular functional category.

Now let's rank them by positional bias. And you get a very different story. The ones that were on the top of the previous one are off the chart here. MCB just barely makes it as number 14. And this A-rich sequence logo, which you might think is something that is all over the place, and in fact, it is.

It has a pretty poor specificity score. It has a high MAP score. Its positional bias is astronomical. It is found in a particular place in many promoters throughout the genome. So this is a way that you quantitate each of these three things, the non-randomness in a learning set, the specificity for that set, and the positional bias within promoters, in general.

So what are the negative controls here? Clusters of size 20, 40, 60, 80, 100 open reading frames, meaning genes for which you might have functional categories. And this allows you to calibrate the false positive rates.

And what you do is you're looking for-- we could use any criteria here. We said that a MAP score might, on average, be 0 if it's random. But if we go up above 10, we'll get a higher enrichment specificity score of 10 to the minus 5 or lower, meaning-- and then we apply these two criteria to the functional categories and to the random controls.

And the functional categories is gratifyingly higher than the random controls. And so we can say that about half of the functional category runs are likely to be real motifs. Of these, about half of those are known. And so the rest are probably new discovered motifs and new discovered regulatons-- regulatory connected genes.

Now, the positive controls, it's said, are harder to come by. There are 29 transcription factors. These are incompletely curated. One of the boons that will come from this systematic analysis of microarray data and functional categories will be a lot of new positive controls. But until we get them, we can't use them. So this is what one can use right now.

And in 21 out of 29 cases, an appropriate motif-- meaning you have to basically rerun AlignACE because you can't really use the weight matrices from the literature. They were derived by slightly different methods. But you can use those to prime AlignACE. It's a trivial thing for AlignACE to now derive a weight matrix. And then you compare it to weight matrices that come out of the tests.

And 21 and 29 work. And of the eight-- the difference between these two is eight-- and of those eight, five were actually an appropriate functional category. So depending on how you interpret these two facts, you can say the false negative rate is 10% to 30%-- not great, but neither the positive control set nor the algorithm are perfect here.

Now, where do we go from here? We need to both generalize and to reduce the assumptions so that we can discover new things. So for example, one of the assumptions we've been making is that motifs act in isolation. We've been discovering motifs one at a time. We'll find the best one. We'll cross it off our list, or we'll filter out subsequence ones. We'll find all the rest.

But what may really be statistically significant, and we may be missing by looking at it one at a time, is motif interactions. And [INAUDIBLE] and coworkers have pursued this with a vengeance. And I think this is a very exciting direction this can go is what proteins-- how two or three or more motifs can interact to produce coregulation.

Then we have these DNA motifs that come out of this microarray data. But what's binding to it? How do we find that connection? Well, one way of many is in vivo crosslinking. There are also so-called one-hybrid acids and so forth. But just think of this conceptually. As you're catching it in the act, you grab it. And then you do proteomics to find which proteins are connected to which nucleic acids.

And the final direction this might be going is we've said that the different columns in the weight matrices are independent. And we've already seen multiple examples in the past fact-- in fact, I emphasize them on purpose-- where the columns are not independent in RNA secondary structure, in CpGs and so forth. And there's some evidence from this paper that the interdependence between columns might be something that you can question.

So in summary, we've talked mainly about clustering and then where you go to check that your clusters are biologically significant, whether you've made discoveries and know the limits of your discoveries. What are the false positive and false negative rates? How do you measure the specificity of your motifs? How do you measure the functional enrichment, things that are casual in the classic literature? So I look forward to seeing you next week. Thank you.