Introduction to Bioinformatics
How to find DNA sequences responsible for gene regulation

Outline:

I. Prelude: Practical pattern recognition
II. Why regulation?
III. How regulation?
IV. Searching for motifs: simple minded approach vs PSSMs
V. How to use PSSMs to find unknown conserved sequences

I. Prelude: Practical pattern recognition

Sitting next to me is our next guest, Giaccomo Fettucini,… is it fair to describe you as the world's foremost connoisseur of Italian pasta?

Well, I can only say that I enjoy my work.

It says here that you're able with a single taste to determine whether a plate of pasta was made by a true Italian chef. Is that right?

It's not as difficult as you make it sound. Anyone could do the same with an appreciation of the elements that make up true Italian pasta.

Hey, I'm anyone. Let's see if you're right. We didn't tell you this, but we arranged for three plates of pasta… Ed, could you bring them in? Up to the challenge, Giaccomo?

I never refuse a plate of good pasta.

Good, let's go. Here's the first… what do you think?

Ah! Delicious! Obviously the work of a master.

Let's see,… you're right! That plate came from *La Belle Noodle*, flown in from Firenze for this show. But how did you know?

Very simple. It has all the markings of a genuine Italian pasta: the red sauce, the hint of garlic, the meatballs that melt in your mouth.

I could do that, if that's all there is to it. Let's try the second plate.

Hmmm. I would place this somewhere in the south of Italy, though there's a hint of oriental influence.

I think we got you this time. That plate came from around the corner at Ming's Yum Yum Café… oh wait a second, I see here that the chef actually is from Naples. That's amazing! But this pasta uses a white sauce, so how could you tell,…

True, the sauce was white, not red, but all the other characteristics were there, so the source was quite obvious.

I get it. A single deviation from your list of requirements is still OK. Well, we have one final plate for you.

Very well… Che Diablo! Take it away!

I have to confess, that plate I made myself. But how did you know? I used a red sauce, added a hint of garlic, and the meatballs…

Yes but you murdered the linguini.

Maybe so, but that's still just one deviation.

I don't mind a different color sauce or some creativity with the spices, but no Italian chef could ever make pasta as limp as this!

Well folks, I hope you caught all that: pasta's Italian if it matches a consensus of characteristics, but one deviation is OK, unless it's in a characteristic that doesn't deviate. I guess that's why we need world famous connoisseurs.

II. Why regulation?

Here we are after only a few thousand years of recorded history, and we now know the secret of life -- DNA. We've figured out the complete genomic sequences of dozens of organisms, including humans, and can predict the amino acid sequences of almost every protein those genomes encode. In principle, though not yet in fact, we can also predict from the sequences of amino acids what functions the proteins will have and even change those functions to suit our wishes.

But don't feel smug: we still don't know how even the simplest living organism is formed.

Upon reflection, this should not surprise you. Suppose I could read every thought in your head, every thought you ever thought, even every thought you haven't thought yet. Everything you were capable of thinking. Would that tell me who you are? Not at all. If every possible thought went through your mind at once, there would be chaos, and you are not chaos.

What's missing is the regulation of your thoughts -- what relationships there are between what is around you and what is called to mind, how one thought connects to another. And that's what's missing from our understanding of genetics at this point: regulation.

At any given moment only a fraction of the genes an organism possesses are expressed as protein, and if they all turned on at the same time... certain death. You have genes that are turned on to protect you when you are overheated, when you are exposed to heavy metals, genes that are expressed only during early embryogenesis, and so forth. To understand how genes determine the form and function of an organism, we must understand not only what genes are but also what regulates their expression.

B. How regulation?

The flow of information from inactive DNA to active protein can be interrupted at any one of several points (Fig. 1). While there are many examples of control at each of the points shown, in most organisms regulation takes place primarily at the first step: the transcription from DNA to RNA. What this means is that if a gene is transcribed, the remaining steps leading to active protein proceed unhindered. Turn on the gene and you turn on the corresponding chemical reaction. So if we understood how transcription is controlled, we'd know a good deal about how a cell controls its capabilities.

SQ1: Why do you think that regulating initiation of transcription is so common as compared, say, to regulating the rate of protein degradation?

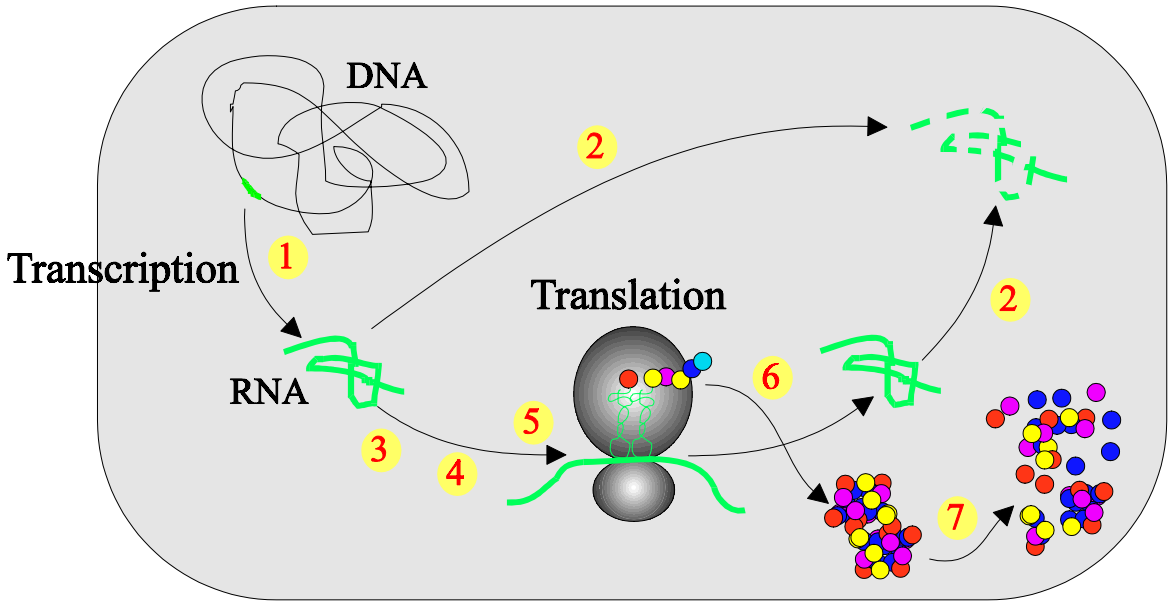

Fig. 1: Control points over gene expression. Choke points in the route from DNA through RNA to active protein (not all shown): **1.** Binding of RNA polymerase/Initiation of transcription, **2**. Degradation of RNA, **3.** Processing of RNA, **4.** Availability of RNA, **5.** Binding of RNA to ribosome/Initiation of translation, **6.** Modification of protein, **7.** Degradation of protein.

We've already looked at one example of transcriptional regulation – remember the *lac* operon a few weeks back? You'll recall that *lac* transcription is regulated by the binding of proteins (CRP, *lac* repressor, RNA polymerase) to specific DNA sequences, and that the binding or not binding of these proteins determined whether transcription of the *lac* operon takes place. Let's look again at transcription, but this time using a cyanobacterial example.

All organisms make metabolic adjustments depending on the environment they find themselves in. For example, cyanobacteria express on set of genes when they are growing on ammonia as a nitrogen source and another when they're deprived of ammonia and forced to use an alternative source. We do not completely understand how cyanobacteria sense nitrogen-deprivation, but some important elements are known (Fig. 2). The protein NtcA responds directly to nitrogen-

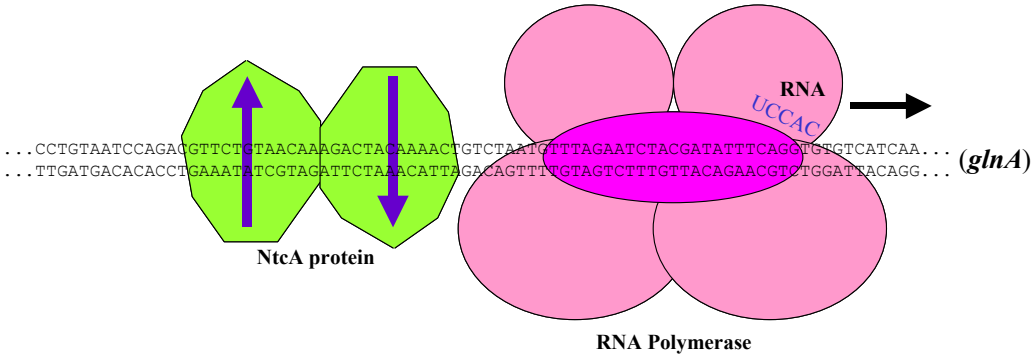

Fig 2. Regulation of expression of *glnA* gene of *Anabaena* PCC 7120. Transcription of *glnA* (encoding glutamine synthetase) requires binding of NtcA protein to a particular sequence upstream from the promoter, the binding site of RNA polymerase. NtcA facilitates binding of RNA polymerase and subsequent transcription.

Strain	gene/operon	Promoter sequence
PCC 7942	*nir* operon	AAAGTT**GTA**GTTTCTGT**TAC**CAATTGCGAATCGAGAACTGCC..**TA**ATC**T**GCCGA**g**
	nirB-ntcB	TTTTTA**GTA**GCAATTGC**TAC**AAGCCTTGACTCTGAAGCCCGC..**T**TAGG**T**GGAGCCAT'
	ntcA	GAAAAA**GTA**GCAGTTGC**TAC**AAGCAGCAGCTAGGCTAGGCCG..**TA**CGG**T**AACG**a**
	glnB	TTGCT**GTA**GCAGTAAC**TAC**AACTGTGGTCTAGTCAGCGGTGT.**TA**CCAAAGAGT**c**
	glnA	TTTTAT**GTA**TCAGCTGT**TAC**AAAAGTGCCGTTTCGGGCTACC..**TA**GGA**T**GAAAG**c**
	amt1	CGAACT**GT**TACATCGAT**TAC**AAAACAACCTTGAGTCTCGCTG..A**A**TGC**T**TACAGAG**a**
PCC 7120	*glnA* (RNAI)	CGTTCT**GTA**ACAAAGAC**TAC**AAAACTGTCTAATGTTTAGAATC.**TA**CGA**T**ATTTC**a**
	nir operon	AATTTT**GTA**GCTACTTA**TAC**TATTTTACCTGAGATCCCGACA..**TA**ACC**T**TAGAAG**t**
	urt operon	AATTTA**GTA**TCAAAAATA**AC**AATTCAATGGTTAAATATCAAAC.**TA**ATA**T**CACAA**t**
	ntcB	AAAGCT**GTA**ACAAAATC**TAC**CAAATTGGGGAGCAAAATCAGC..**TA**ACT**T**AATTGA**a**
	devBCA	TCATTT**GTA**CAGTCTGT**TAC**CTTTACCTGAAACAGATGAATG..**TA**GAA**T**TTAT**a**
PCC 6803	*amt1*	TGAAAA**GTA**GTAAATCA**TAC**AGAAAACAATCATGTAAAAA....**T**TGAA**T**ACTCT**aa**
	glnA	AAAATG**GTA**GCGAAAAA**TAC**ATTTTCTAACTACTTGACTCTT..**TA**CGA**T**GGATAGT**c**
	glnB	CAAACG**GTA**CTGATTTT**TAC**AAAAAAACTTTTGGAGAACATGT.**TA**AAAGTGTCT**gg**
	icd	AATTTC**GTA**ACAGCCAA**T**G**C**AATCAGAGCCTCCAGAAAGGAT..**TA**TGA**T**CTGCTCC**g**
	rpoD2-V	AAGTTT**GTA**TCACGAAT**TAC**ACTGCCGTGAAAATTTAACGA...**TA**TTT**T**GGACA**g**
PCC 7601	*glnA* (P1)	GAATCT**GTA**ACAAAGAC**TAC**AAAAATTCTTAATGTCATATCCT.**TA**GGA**T**ATTCCAG**g**
PCC 6903	*glnN*	TTTTTT**GT**GCGCGTTTA**TAC**CAATCAAGTGCGATCTAATCGG..**TA**TCT**T**TTTTAT**c**
PCC 7002	*nrtP*	TAAAGA**GTA**TCAGCGGT**TAC**GAATTTAGCGAAGAAAGAATGTGA**T**TCTT**T**ATCAC**a**
WH 7803	*ntcA*	GGAACC**GT**GTGCGTTGC**TAC**AGGGTGGGAATCGATCGCTCCT..**TA**ATT**T**CCTTGA**a**
		GTA ..(8).. **TAC** ..(20-24).. **TA**..(3)..**T**
		Consensus NtcA binding site **promoter (-10)**

Fig. 3: Alignment of known NtcA binding sites upstream from cyanobacterial genes regulated by nitrogen deprivation. The accepted consensus binding sequence is given below. The organisms are: *Synechococcus* PCC 7942, *Anabaena* PCC 7120, *Synechocystis* PCC 6803, *Tolypothrix* PCC 7601, *Pseudanabaena* PCC 6903, *Synechococcus* PCC 7002, and *Synechococcus* WH 7803. Taken from Herrero et al [J Bacteriol (2001) 183:411-425.

deprivation, changing its conformation so that it becomes able to bind to specific sequences upstream from nitrogen-regulated genes. Many sites recognized by NtcA protein have been determined by cutting DNA to which NtcA has been bound and determining what sequence NtcA protects. Some sites are shown in Fig. 3

Don't be fooled into thinking the problem of how nitrogen regulates gene expression has been solved! There are far more genes regulated by nitrogen than shown in the table. How can we find out what they are? One way is to repeat the NtcA-binding experiments with all sequences upstream from genes. With as many as 8000 genes in a cyanobacterium, this is far from practical! An alternative approach made possibly by the availability of genomic sequences is to look computationally for sites that may bind NtcA.

How to predict protein binding sites? A simple minded approach would be to take the consensus sequence (at the bottom of Fig. 3) and search for that sequence throughout the genome of a cyanobacterium. Let's try it:

SQ2. Find all sequences in the genome of *Anabaena* PCC 7120 that match the consensus sequence. To do this use the following BioLingua forms:

```
(DEFINE seq AS (SEQUENCE-OF (CHROMOSOME-OF A7120)) DISPLAY off)
(PATTERN-CAPTURE-ALL "GTA.{8}TAC.{20,24}TA...T" seq)
(LENGTH *)
```

(Needless to say, you'll have to execute these forms one at a time). The first form puts the sequence of the *Anabaena* chromosome into the variable `seq`. The second

form finds and captures all instances in the chromosome of the given pattern. The pattern is read this way:

1. **Starts with** `GTA`
2. **Followed by exactly 8 characters of any type (. matches anything)**
3. **Followed by** `TAC`
4. **Followed by anywhere from 20 to 24 characters of any type**
5. **Followed by** `TA`**, then three characters, then** `T`

The exercise above illustrates one shortcoming with the approach: There are too many sequences that match the consensus pattern. There aren't that many genes regulated by NtcA! A second shortcoming can be appreciated by reinspecting Fig. 3:

SQ3. How many of the proven NtcA-binding sites shown in Fig. 3 match the consensus pattern?

So we have a problem: Finding sequences by matching the consensus pattern gives too many false positives and too many false negatives. We need another approach.

II. Searching for motifs: simple minded approach vs PSSMs

The dialog in **Section I** hints at the problem. An expert would not apply a strict consensus sequence, or apply a strict rule (e.g. one mismatch allowed) but instead would consider a sequence in light of his accumulated experience. He would look at many characteristics, perhaps some subconsciously, and allow candidates the same kinds of imperfections as he has observed with real sequences, but only those kinds.

The ultimate expert is NtcA itself. Short of an in depth interview with a cooperative protein, the best we can do is to try to extrapolate from our own experience. Here's an analogous situation. Suppose you want to find all ways that people spell the word "color". You might look for all words that differed from only one letter, e.g. "coler", "color", "kolor". Unfortunately, this procedure would also give you "polor" and "colox", which are not likely spelling errors. If you wanted to limit your set to those instances where people *mean* color, then you could collect a training set of words where by context you're convinced the intent was "color" and see what kinds of mistakes were made. You'd probably find that the vowels showed some variability but the consonants were seldom missed. Learning from this, you might accept a word even with two errors (e.g. culer) but not one that replaced "l" with some other consonant.

A part of this expert process can be captured by what are called position-specific scoring matrices (PSSMs). Given an aligned set of sequences, it is very easy to construct a PSSM. Let's consider again the sequences surrounding the proven NtcA binding sites in *Nostoc* (Table 1A). The consensus sequence used only the six most highly conserved nucleotides within the NtcA-binding site: GTA...(N_8)...TAC. Ignoring the other positions tosses out a good deal of potentially useful information, as can be seen from the table of occurrences (Table 1B) and the PSSM derived from it (Table 1C). The latter is taken directly from the former by dividing the number of occurrences by the total number of sequences.

The PSSM gives us a tool to score how close any sequence is to the collected sequences used to create the scoring matrix (also called the training sequences). You would expect that a sequence close to the training sequences would tend to have higher scores at each position. The total score, i.e. the product of the scores at each position, should be higher that of most other sequences of

Table 1: Examples of position-specific scoring matrices from sequence alignment

A. Sequence alignment[a]

urt-71	A	T	T	T	A	G	T	A	T	C	A	A	A	A	A	T	A	A	C	A	A	T	T	C
glnA-71	G	T	T	C	T	G	T	A	A	C	A	A	A	G	A	C	T	A	C	A	A	A	A	C
nirA-71	A	T	T	T	T	G	T	A	G	C	T	A	C	T	T	A	T	A	C	T	A	T	T	T
ntcB-71	A	A	G	C	T	G	T	A	A	C	A	A	A	A	T	C	T	A	C	C	A	A	A	T
devBCA-71	C	A	T	T	T	G	T	A	C	A	G	T	C	T	G	T	T	A	C	C	T	T	T	A

B. Table of occurrences[a]

A	3	2	0	0	1	0	0	5	2	1	3	4	3	2	2	1	1	5	0	2	4	2	2	1
C	1	0	0	2	0	0	0	0	1	4	0	0	2	0	0	2	0	0	5	2	0	0	0	2
G	1	0	1	0	0	5	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0
T	0	3	4	3	4	0	5	0	1	0	1	1	0	2	2	2	4	0	0	1	1	3	3	2

C. Position-specific scoring matrix (B = 0)[b]

A	.60	.40	0	0	.20	0	0	.80	.40	.20	.60	.80	.60	.40	.40	.20	.20	.80	0	.40	.80	.40	.40	.20
C	.20	0	0	.40	0	0	0	0	.20	.80	0	0	.40	0	0	.40	0	0	.80	.40	0	0	0	.40
G	.20	0	.20	0	0	.80	0	0	.20	0	.20	0	0	.20	.20	0	0	0	0	0	0	0	0	0
T	0	.60	.80	.60	.80	0	.80	0	.20	0	.20	.20	0	.40	.40	.40	.80	0	0	.20	.20	.60	.60	.40

D. Position-specific scoring matrix (B = √N = 2.2)[c]

A	.51	.38	.099	.099	.24	.099	.099	.79	.38	.24	.51	.65	.51	.38	.38	.24	.24	.79	.099	.38	.65	.38	.38	.24
C	.19	.056	.056	.33	.056	.056	.056	.056	.19	.61	.056	.056	.33	.056	.056	.33	.056	.056	.75	.33	.056	.056	.056	.33
G	.19	.056	.19	.056	.056	.75	.056	.056	.19	.056	.19	.056	.056	.19	.19	.056	.056	.056	.056	.056	.056	.056	.056	.056
T	.099	.51	.65	.51	.65	.099	.79	.099	.24	.099	.24	.24	.099	.38	.38	.38	.65	.099	.099	.24	.24	.51	.51	.38

E. Position-specific scoring matrix (B = 0.1)[c]

A	.59	.40	.006	.006	.20	.006	.006	.99	.40	.20	.59	.79	.59	.40	.40	.20	.20	.99	.006	.40	.79	.40	.40	.20
C	.20	.004	.004	.40	.004	.004	.004	.004	.20	.79	.004	.004	.40	.004	.004	.40	.004	.004	.98	.40	.004	.004	.004	.40
G	.20	.004	.20	.004	.004	.98	.004	.004	.20	.004	.20	.004	.004	.20	.20	.004	.004	.004	.004	.004	.004	.004	.004	.004
T	.006	.59	.79	.59	.79	.006	.99	.006	.20	.006	.20	.20	.006	.40	.40	.40	.79	.006	.006	.20	.20	.59	.59	.40

F. Position-specific scoring matrix: Log-odds form (B = 0.1)[c,d]

A	0.2	0.4	2.2	2.2	0.7	2.2	2.2	0.0	0.4	0.7	0.2	0.1	0.2	0.4	0.4	0.7	0.7	0.0	2.2	0.4	0.1	0.4	0.4	0.7
C	0.7	2.5	2.5	0.4	2.5	2.5	2.5	2.5	0.7	0.1	2.5	2.5	0.4	2.5	2.5	0.4	2.5	2.5	0.0	0.4	2.5	2.5	2.5	0.4
G	0.7	2.5	0.7	2.5	2.5	0.0	2.5	2.5	0.7	2.5	0.7	2.5	2.5	0.7	0.7	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
T	2.2	0.2	0.1	0.2	0.1	2.2	0.0	2.2	0.7	2.2	0.7	0.7	2.2	0.4	0.4	0.4	0.1	2.2	2.2	0.7	0.7	0.2	0.2	0.4

[a] Alignment of proven *Anabaena* NtcA-binding sites, as shown in Figure 3. Boxes shaded in red are the positions of the accepted consensus sequence.

[b] Shading indicates fraction of occurances for that base at that position: red (1.0), orange (0.8), yellow (0.6).

[c] The background frequencies used to calculate the scores are **A** = **T** = 0.32; **C** = **G** = 0.18. These are the observed average nucleotide frequencies in intergenic sequences of *Nostoc* PCC 7120. Table 1D was calculated with the default scoring system used by the Gibbs Sampler, and Table 1E used the default scoring system of Meme.

[d] Each element of the table is equal to the negative $\log_{10}$ of the corresponding element of Table 1E.

Table 2: Example of scoring a sequence with a PSSM

urt-71	A	T	T	T	A	G	T	A	T	C	A	A	A	A	A	T	A	A	C	A	A	T	T	C
Score[a]	.60	.60	.80	.60	.20	1.0	1.0	1.0	.20	.80	.60	.80	.60	.40	.40	.40	.20	1.0	1.0	.40	.80	.60	.60	.40
w/ps'counts[b]	.51	.51	.65	.51	.24	.75	.79	.79	.24	.61	.51	.65	.51	.38	.38	.38	.24	.79	.75	.38	.65	.51	.51	.33
Normal'd[c]	1.6	1.6	2.0	1.6	.75	4.2	2.5	2.5	.75	3.4	1.6	2.0	1.6	1.2	1.2	1.2	.75	2.5	4.2	1.2	2.0	1.6	1.6	1.8

[a] Scoring matrix from Table 1C used. The product of the elemental scores is 9×10^{-7}.

[b] Scoring matrix from Table 1D used.

[c] Scoring matrix from Table 1D used, correcting for background nucleotide frequencies by dividing the raw score (with pseudocounts) by the frequency of the given nucleotide. The product of the elemental scores is 3.2×10^{5}.

similar length. Table 2 shows an example of how the PSSM would be used to score the sequence ATTTAGTATCAAAAATAACAATTC. The score of $9x10^{-7}$ does not have any meaning except in comparison with scores of other sequences of the same length, calculated using the same scoring table.

SQ4: What (on the basis of this small training set) would seem to be the most informative columns in predicting whether a sequence is an NtcA binding site?

SQ5: What does ".60" in the upper left corner of Table 1C mean?

SQ6: How was the score $9x10^{-7}$ in Table 2 obtained?

For now, don't worry about the adjustments to the PSSM shown in Figs. 1D-F. We'll discuss them in class.

IV. How to use PSSMs to find unknown conserved sequences

In the present case we have the advantage of knowing already in some cases where NtcA binds. More often, we have only a collection of genes that are possibly coregulated – may they turn on at the same time and place during development or turn on in response to the same environmental stimulus (e.g. heat, hormone, etc.). From physiological and genetic information, we might collect genes whose upstream regions should share some regulatory sequence in common. But how to find that sequence? How can we build a PSSM when we have no idea how to align the very different upstream regions?

BioLingua offers a tool that does just that: Meme. To use it, you need to supply a training set of sequences that you believe may have a common sequence. For example, the training set might consist of all genes you believe are regulated by iron. Or, it might consist of all orthologs of a single gene, for example orthologs of *glnA* in all available cyanobacteria. Either way, you give Meme a set of sequence, then it cranks away at it looking for motifs within the sequences that appear more frequently than you would expect by chance.