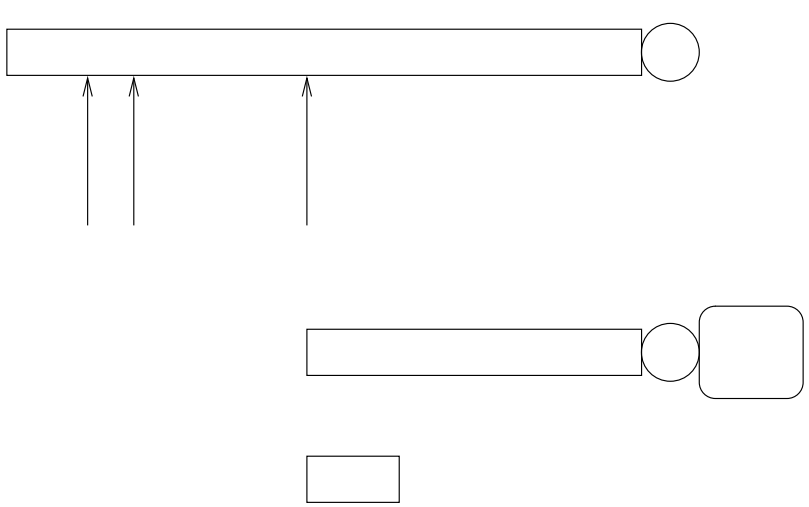
Homework Assignment — Due Thursday

1. Read the paper on SAGEmap.
2. At the SAGEmap Web site, find genes that are over- or under-expressed by a factor of at least 5 in some brain tumor as compared to some library (or libraries) or normal brain tissue.
3. Taking one of the genes that is overexpressed in a certain brain tumor, find out what tissues normally express that gene.

SAGE — Serial Analysis of Gene Expression

1. Create cDNA sequences from poly−A+ RNA with biotinylated ends
2. Cut at every CATG (Nla3 site)
3. Isolate 3'−most fragments using streptavidin beads
4. Chop off the CATG plus next 10 nucleotides (SAGE tags)
5. Sequence the tags



SAGE Output

For a given cDNA library, e.g., from a certain tissue or tumor, we get a list of how many times each of the possible 4^{10} 10-mer tags was found:

AAAAAAAAAA	14
AAAAAAAAAC	1074
AAAAAAAAAG	0
...	
TTTTTTTTTT	192

Frequency of Erroneous SAGE Tags

Typically, SAGE tags are generated by single-pass sequencing, i.e., DNA isn't re-sequenced to improve accuracy. Hence, about 1%-3% of the nucleotides are determined incorrectly. This means that about 1 of 10 (at 1% error) or 1 of 4 (at 3% error) tags is erroneous.

Suppose that a given tag, say ACCGTATAGC, is not the tag for a real gene that is expressed in the cell type under investigation. How many bogus copies of that tag can we expect to arise from sequencing errors?

Generally, A Given Bogus Tag Will Be Rare

Suppose 100,000 tags are sequenced. Assuming 1%/bp sequencing errors, about 10,000 of these tags will be incorrectly sequenced. Since there are about 1 million tags, the average number of times that each of them is incorrectly observed is about 10,000/1,000,000 = 0.01.

Thus, if we observe the tag ACCGTATAGC 2 times, then it is probably real. Right?

A Closer Look

To see that it is worthwhile to look more closely at the number of times a tag can be erroneously observed, suppose the tag ACCGTATAGC is observed 2 times. Suppose also that we observe CCCGTATAGC (different in the first letter) 2000 times. We expect the leading C to be incorrectly sequenced about 20 times and for it to be read as A about 7 times. Thus, the surprising thing is that ACCGTATAGC is only observed twice!

Neighbors of a Putative SAGE Tag

Consider a tag T, say T = ACCGTATAGC. For now, let's consider only tags with a single sequencing error. We'll call a sequence S a *neighbor* of T if S can be turned into T with one sequencing error.

Given T, we're going to look at each of its neigbors, S, use the number of observations of S to estimate how often it actually occurs, estimate the number of times that a sequencing error in S gave us a T, and add these values over all neighbors of T.

If You Like Equations...

T is a 10-mer.

$N(T)$ is the set of neighbors of T.

Assume the probability of a tag being sequenced incorrectly is 10%.

$count(S)$ is the number of times S is observed in our SAGE data.

$V(T)$ is the expected contribution of tags of neighbors of T to observations of T.

Then $V(T) = \sum_{S \in N(T)} \frac{0.1 \times count(S)}{\#N(S)}$. Here, $\#N(S)$ is the number of things in $N(S)$.

Evaluating $V(T)$

$V(T) = \sum_{S \in N(T)} \frac{0.1 \times count(S)}{\#N(S)}$.

How many neighbors does T have? What are they?

Counting Neighbors of a 10-Mer

A 10-mer can be changed into:

- 30 different 10-mers by a single substitution
- at most 10 different 9-mers by a deletion (note: AAAAAAAAAA can be changed into just one unique 9-mer by deletion)
- at most 44 different 11-mers by an insertion (note: with AAAAAAAAAA, there are 11 insertions that give an all-A 11-mer)

More Ideas

To estimate how many observations of T will be contributed by one of its neighbors, say S, we can look at the quality value (e.g., from the `phred` program) for the relevant position in S.

We can deal with simultaneous effects on T and S (sequencing errors in T can give S, and vice versa), trying to predict the actual counts from the observed counts. This leads to a (large, sparse, diagonally dominant) system of linear equations.

Picking Up Where We Left Off on Thursday...

At the end of the previous lecture, I was left with 617 gene-mRNAcount pairs for housekeeping genes.

AAMP	14
ACADVL	71
ACTN4	68
. . .	
ALDOA	658
. . .	

One Use of SAGE Data and Genomic Sequences

Hypothesis: Many housekeeping genes have primitive promoters, identified by a CpG island, that do not provide accurate control of the expression level. Instead, their level of expression is controlled by factors including length of the gene (short genes are more highly expressed) and proximity to other genes (close genes result in transcriptional interference, reducing expression).

Experiment

For each of the 617 housekeeping genes where we have SAGE data for their expression level, we'll determine the position in the human genome if possible. Then, we'll know the gene's length and be able to predict the distance to the closest neighboring gene. Applying statistical techniques will test our hypothesis by looking for correlations between mRNA count and either gene length or distance to the closest neighboring gene.

How many known human genes have a known genomic location with accurate sequence data?

Placing Human Genes on the Genome

Of the 617 housekeeping genes, 547 had a predicted position in the August 2001 “freeze” at the Genome Browser (Univ. Calif. Santa Cruz). Indeed, 86 had more then one predicted location. Many of these cases were putative alternate splicings, but many were genuine duplications. Given a choice, our program prefered the prediction with more exons.

Manual curation moved 7 of these predictions to other locations. In 11 cases, we discarded all prediction because none looked adequate. This left 536 genes.

Extending the Predictions

The original predicted locations were based on RefSeqs, which are NCBI's curated versions of genes. However, it appeared that frequently the non-translated portions of the gene are not completely determined. In particular, we found 141 cases where gene predictions from ENSEMBL (another major provider of genome annotation) predicted longer genes based typically on EST information.

More Manual Curation

Most of the 141 cases where ENSEMBL prediction extended the RefSeq involved a small number of nucleotides. We checked cases where the extension exceeded 200 bp. In 25 of them, it appeared that the ENSEMBL prediction was too long, and for those genes we returned to the RefSeq prediction.

Also, there were 5 cases where the ENSEMBL prediction added one or more exons to the RefSeq prediction. Since accurate exon counts are important in our analyses we simply discarded the 5 genes.

Looking for CpG Islands.

All or nearly all of the human housekeeeping genes are said to have a CpG island around the transcription start site. (A CpG island is a region having an unusally high percentage of CG dinucleotides.) We checked the 531 genes and found that 25 lacked a CpG island.

Manual inspection of those 25 indicated problems in almost every case. Some appeared to be pseudo-genes, and in other cases there was EST or mRNA evidence that the RefSeq and ENSEMBL predictions were both wrong. We discarded the most troublesome cases and manually corrected the others, leaving 512 genes.