

# Better than Chance: the importance of null models

[http://users.soe.ucsc.edu/~karplus/papers/  
better-than-chance-sep-11.pdf](http://users.soe.ucsc.edu/~karplus/papers/better-than-chance-sep-11.pdf)

Kevin Karplus  
karplus@soe.ucsc.edu

Biomolecular Engineering Department  
University of California, Santa Cruz

13 Sept 2011



# Outline of Talk

- 🐞 What is a protein?
- 🐞 The folding problem and variants on it.
- 🐞 What is a null model (or null hypothesis) for?
- 🐞 Example 1: is a conserved ORF a protein?
- 🐞 Example 2: is residue-residue contact prediction better than chance?
- 🐞 Example 3: how should we remove composition biases in HMM searches?



# What is a protein?

- 🦨 There are many abstractions of a protein: a band on a gel, a string of letters, a mass spectrum, a set of 3D coordinates of atoms, a point in an interaction graph, . . . .
- 🦨 For us, a protein is a long skinny molecule (like a string of letter beads) that folds up consistently into a particular intricate shape.
- 🦨 The individual “beads” are amino acids, which have 6 atoms the same in each “bead” (the *backbone* atoms: N, H, CA, HA, C, O).
- 🦨 The final shape is different for different proteins and is essential to the function.
- 🦨 The protein shapes are important, but are expensive to determine experimentally.



# Folding Problem

The *Folding Problem*:

If we are given a sequence of amino acids (the letters on a string of beads), can we predict how it folds up in 3-space?

---

MTMSRRNTDA ITIHSILDWI EDNLESPLSL EKVSERSGYS KWHLQRMFKK  
ETGHSLGQYI RSRKMTEIAQ KLKESNEPIL YLAERYGFES QQTLTRTFKN  
YFDVPPHKYR MTNMQGESRF LHPLNHYNS



Too hard!



# Fold-recognition problem

The *Fold-recognition Problem*:

Given a sequence of amino acids  $A$  (the *target* sequence) and a library of proteins with known 3-D structures (the *template* library),

figure out which templates  $A$  match best, and align the target to the templates.

- 👉 The backbone for the target sequence is predicted to be very similar to the backbone of the chosen template.
- 👉 Progress has been made on this problem, but we can usefully simplify further.



# Remote-homology Problem

The *Homology Problem*:

Given a target sequence of amino acids and a library of protein *sequences*, figure out which sequences  $A$  is similar to and align them to  $A$ .

- 🦖 No structure information is used, just sequence information. This makes the problem easier, but the results aren't as good.
- 🦖 This problem is fairly easy for recently diverged, very similar sequences, but difficult for more remote relationships.



# Scoring (Bayesian view)

- 👉 A model  $M$  is a computable function that assigns a probability  $P(A | M)$  to each sequence  $A$ .
- 👉 When given a sequence  $A$ , we want to know how likely the model is. That is, we want to compute something like  $P(M | A)$ .
- 👉 Bayes Rule:

$$P(M | A) = P(A | M) \frac{P(M)}{P(A)} .$$

- 👉 Problem:  $P(A)$  and  $P(M)$  are inherently unknowable.



# Null models

Standard solution: ask how much more likely  $M$  is than some *null hypothesis* (represented by a *null model*  $N$ ):

$$\frac{P(M|A)}{P(N|A)} = \frac{P(A|M)}{P(A|N)} \frac{P(M)}{P(N)}.$$

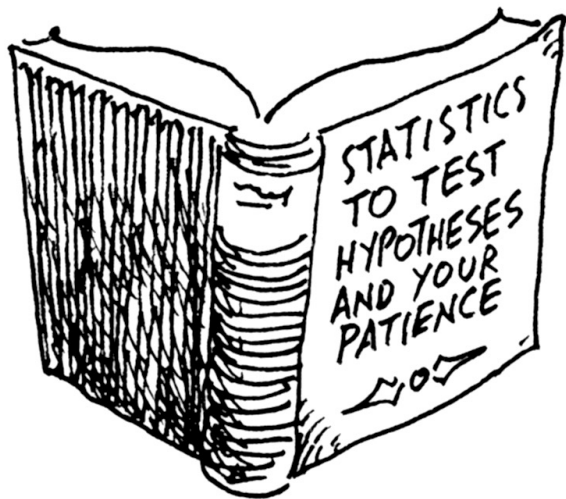
↑                      ↑                      ↑

posterior odds                      likelihood ratio                      prior odds





# Test your hypothesis



Thanks to Larry Gonick *The Cartoon Guide to Statistics*

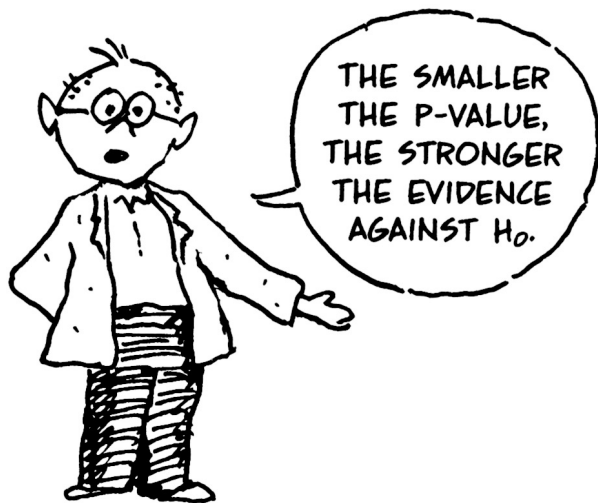


# Scoring (frequentist view)

- 🐘 We believe in models when they give a large score to our observed data.
- 🐘 Statistical tests (p-values or E-values) quantify how often we should expect to see such good scores “by chance”.
- 🐘 These tests are based on a null model or null hypothesis.



# Small p-value to reject null hypothesis



Thanks to Larry Gonick *The Cartoon Guide to Statistics*



# Statistical Significance (2 approaches)

**Markov's inequality** For any scoring scheme that uses

$$\ln \frac{P(\text{seq} | M)}{P(\text{seq} | N)}$$

the probability of a score better than  $T$  is less than  $e^{-T}$  for sequences distributed according to  $N$ .

**Parameter fitting** For “random” sequences drawn from some distribution other than  $N$ , we can fit a parameterized family of distributions to scores from a random sample, then compute  $P$  and  $E$  values.



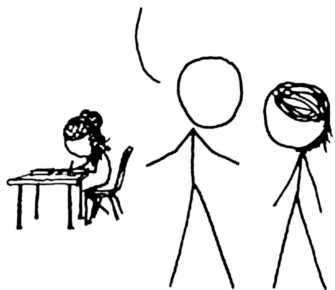
# Null models

- 🐞 P-values (and E-values) often tell us nothing about how good our hypothesis is.
- 🐞 What they tell us is how bad our null model (null hypothesis) is at explaining the data.
- 🐞 A badly chosen null model can make a very wrong hypothesis look good.



I CAN'T BELIEVE SCHOOLS  
ARE STILL TEACHING KIDS  
ABOUT THE NULL HYPOTHESIS.

I REMEMBER READING A BIG  
STUDY THAT CONCLUSIVELY  
DISPROVED IT *YEARS* AGO.



<http://xkcd.com/892/>



## Example 1: long ORF

- 🧑 A colleague found an ORF in an archæal genome that was 388 codons long and was wondering if it coded for a protein and what the protein's structure was.
- 🧑 We know that short ORFs can appear “by chance”.
- 🧑 So how likely is this ORF to be a chance event?



# Null Model 1a: no selection

- 🧪 G+C content bias. (GC is 35.79%, AT is 64.21%.)
- 🧪 Probability of start codon  
 $ATG = 0.321 * 0.321 * 0.179 = 0.01845$
- 🧪 Probability of stop codon  
TAG= 0.1845, TGA=0.01845, TAA=0.0331, so  $p(\text{STOP})=0.06999$
- 🧪  $P(\text{ATG}, 387 \text{ codons without stop}) =$   
 $p(\text{ATG})(1 - p(\text{STOP}))^{387} = 1.18e - 14$
- 🧪 E-value in double-strand genome (6e6 bases)  $\approx 7.05e - 08.$
- 🧪 We can easily reject this null hypothesis!





# Null Model 1b: codon (3-mer) bias

- 👉 Count 3-mers in double-stranded genome.
- 👉 Probability of ATG start codon: 0.01567
- 👉 Probability of stop codon: 0.07048
- 👉  $P(\text{ATG, 387 codons without stop}) = p(\text{ATG})(1 - p(\text{STOP}))^{387} = 8.15e - 15$
- 👉 E-value in genome  $\approx 4.87e - 08$ .
- 👉 We can easily reject this null hypothesis!

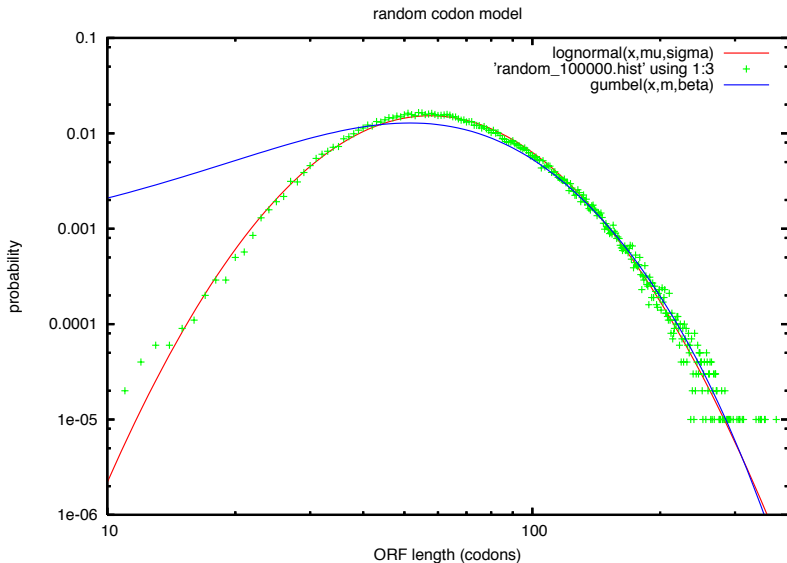


## Null Model 2: reverse of gene

- 🐞 ORF is on the opposite strand of a known 560-codon thermosome gene!
- 🐞 What is the probability of this long an ORF, on opposite strand of known gene?
- 🐞 Generative model: simulate random codons using the codon bias of the organism, take reverse complement, and see how often ORFs 388-long or longer appear.
- 🐞 Taking 100,000 samples, we get estimates of P-value  $\approx 1.5e-05$
- 🐞  $\approx 3000$  genes, giving us an E-value  $\approx 0.045$
- 🐞 Hard to reject null!

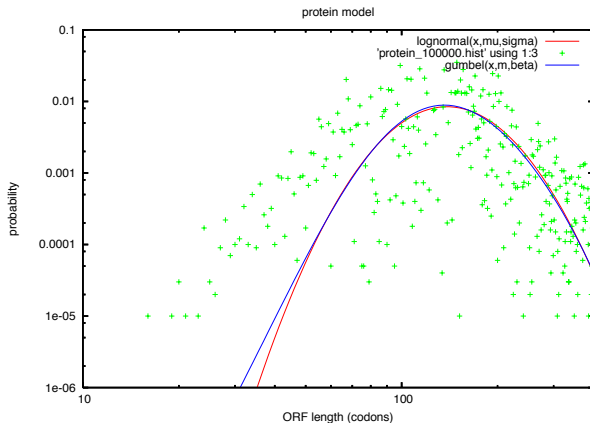


# Null Model 2 histogram



# Null Model 3

- 🐉 Same sort of simulation, but use codons that code for the right protein on the forward strand.
- 🐉 P-value and E-value  $\approx 0.0025$  for long ORFs on the reverse strand of genes coding for this protein.



# Protein or chance ORF?



Thanks to Larry Gonick *The Cartoon Guide to Statistics*



# Not a protein

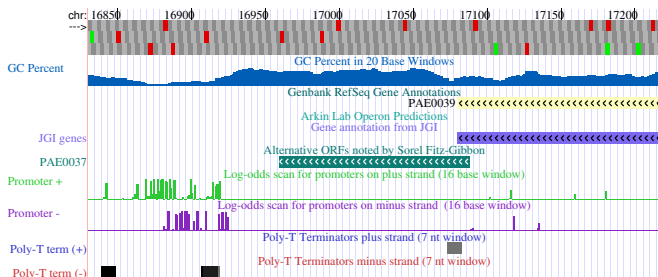
- + A tblastn search with the sequence revealed similar ORFs in many genomes.
- All are on opposite strand of homologs of same gene.
- “Homologs” found by tblastn often include stop codons.
- There is no evidence for a TATA box upstream of the ORF.
- No strong evidence for selection beyond that explained by known gene.

**Conclusion:** it is rather unlikely that this ORF encodes a protein.



# Example 1b: another ORF

🦘 pae0037: ORF, but probably not protein gene in *Pyrobaculum aerophilum*



🦘 Promoter on wrong side of ORF.

🦘 High GC content (need local, not global, null)

🦘 Strong RNA secondary structure.



## Example 2: contacts

- 🐞 Is residue-residue contact prediction better than chance?
- 🐞 Early predictors (1994) reported results that were 1.4 to 5.1 times “better than chance” on a sample of 11 proteins.
- 🐞 But they used a uniform null model:

$$P(\text{residue } i \text{ contacts residue } j) = \text{constant} .$$

- 🐞 A better null model:

$$P(\text{residue } i \text{ contacts residue } j) = \\ P(\text{contact} \mid \text{separation} = |i - j|) .$$

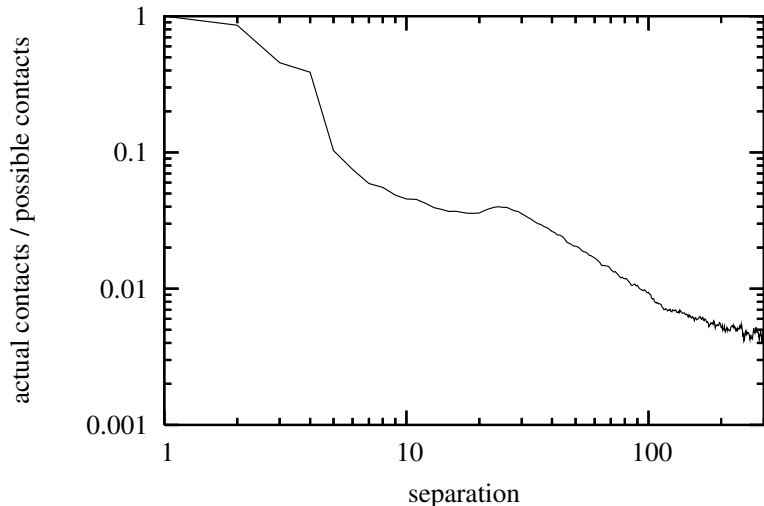




# P(contact|separation)

Using CASP definition of contact, CB within 8 Å, CA for GLY.

dunbrack-40pc-3157-CB8



# Can get accuracy of 100%

- 👉 By ignoring chain separations, the early predictors got what sounded like good accuracy (0.37–0.68 for L/5 predicted contacts)
- 👉 But just predicting that  $i$  and  $i + 1$  are in contact would have gotten accuracy of 1.0 for even more predictions.
- 👉 More recent work has excluded small-separation pairs, with different authors choosing different thresholds.
- 👉 CASP uses separation  $\geq 6$ ,  $\geq 12$ , and  $\geq 24$ , with most focus on  $\geq 24$ .



# Separation as predictor


If we predict all pairs with given separation as in contact, we do much better than uniform model.

sep	$P(\text{contact} \mid  i-j  = \text{sep})$	$P(\text{contact} \mid  i-j  \geq \text{sep})$	"better than chance"
6	0.0751	0.0147	4.96
9	0.0486	0.0142	3.42
12	0.0424	0.0136	3.13
24	0.0400	0.0116	3.46



# Evaluating contact prediction

Two measures of contact prediction:

 Accuracy:

$$\frac{\sum \chi(i, j)}{\sum 1}$$

 Weighted accuracy:

$$\frac{\sum \chi(i, j) / P(\text{contact} \mid \text{separation} = |i - j|)}{\sum 1}$$

= 1 if predictions no better than chance, independent of separations for predicted pairs.



# CASP7 Contact prediction

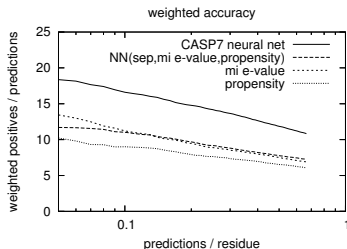
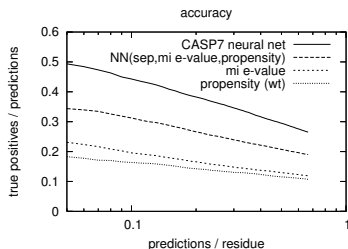
- 🐘 Use mutual information between columns of thinned alignment ( $\leq 50\%$  identity)
- 🐘 Compute e-value for mutual info (correcting for small-sample effects).
- 🐘 Compute rank of e-value within protein.
- 🐘 Feed  $\log(\text{e-value})$ ,  $\log(\text{rank})$ , contact potential, joint entropy, and separation along chain for pair, and amino-acid profile, predicted burial, and predicted secondary structure for window around each residue of pair into a neural net.



# Now doing better

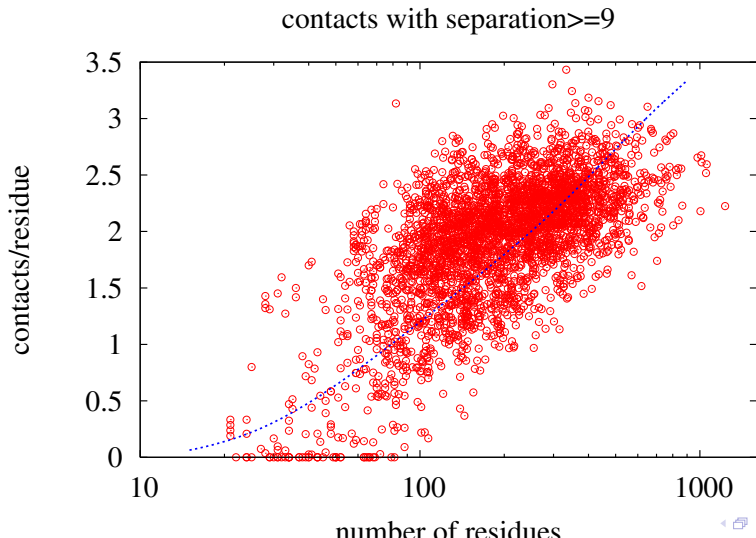
separation  $\geq 9$

Predictions/residue taken separately for each protein.



# Contacts per residue

We can also use our null model to predict the number of contacts per residue (which is not a constant).



## Example 3: HMM

- 👉 Hidden Markov models assign a probability to each sequence in a protein family.
- 👉 A common task is to choose which of several protein families (represented by different HMMs) a protein belongs to.





# Standard Null Model

🐘 Null model is an i.i.d (independent, identically distributed) model.

$$P(A \mid N, \text{len}(A)) = \prod_{i=1}^{\text{len}(A)} P(A_i) .$$

$$P(A \mid N) = P(\text{sequence of length } \text{len}(A)) \\ \prod_{i=1}^{\text{len}(A)} P(A_i) .$$



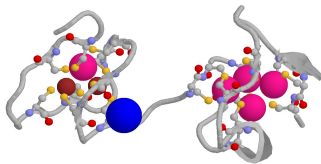
# Composition as source of error

- 🐞 When using the standard null model, certain sequences and HMMs have anomalous behavior. Many of the problems are due to unusual composition—a large number of some usually rare amino acid.
- 🐞 For example, metallothionein, with 24 cysteines in only 61 total amino acids, scores well on any model with multiple highly conserved cysteines.

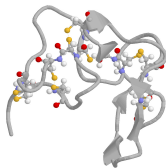


# Composition examples

Metallothionein Isoform II (4mt2)

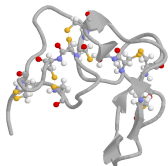


Kistrin (1kst)

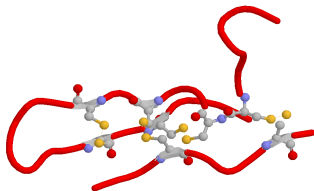


# Composition examples

Kistrin (1kst)



Trypsin-binding domain of Bowman-Birk Inhibitor (1tabl)



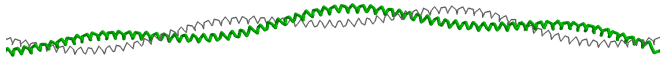
# Reversed model for null

- 🧑 We avoid this (and several other problems) by using a reversed model  $M^r$  as the null model.
- 🧑 The probability of a sequence in  $M^r$  is exactly the same as the probability of the reversal of the sequence given  $M$ .
- 🧑 This method corrects for composition biases, length biases, and several subtler biases.



# Helix examples

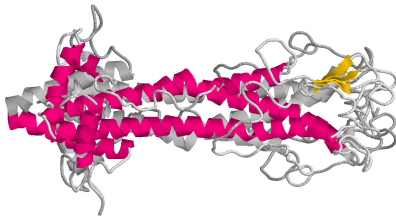
Tropomyosin (2tmaA)



Colicin Ia (1cii)

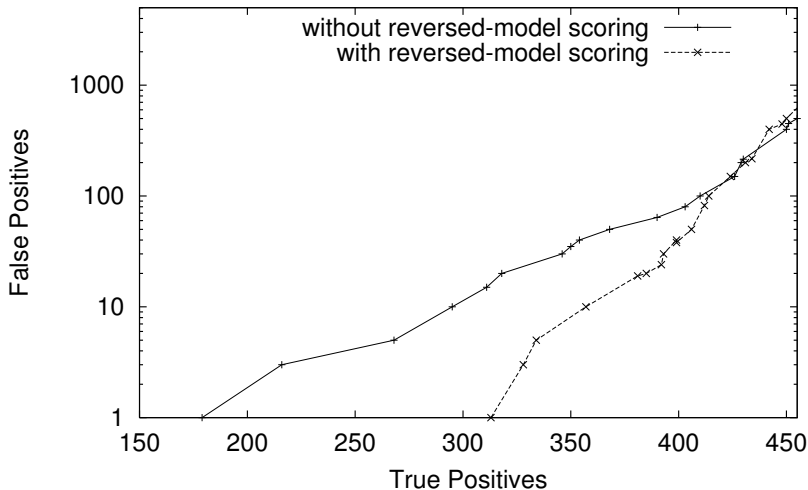


Flavodoxin mutant (1vsgA)

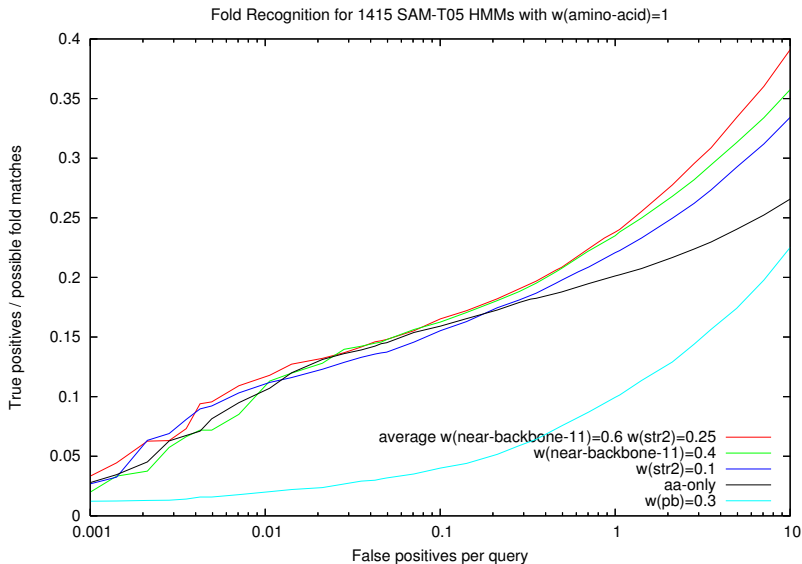


# Improvement from reversed model

SCOP whole chains



# Fold recognition results





# Take-home messages

- 🐘 Base your null models on biologically meaningful null hypotheses, not just computationally convenient math.
- 🐘 Generative models and simulation can be useful for more complicated models.
- 🐘 Picking the right model remains more art than science.



# Web sites

**List of my papers:** <http://users.soe.ucsc.edu/~karplus/papers/paper-list.html>

**These slides:** <http://users.soe.ucsc.edu/~karplus/papers/better-than-chance-sep-11.pdf>

**Reverse-sequence null:** Calibrating E-values for hidden Markov models with reverse-sequence null models.  
*Bioinformatics*, 2005. 21(22):4107–4115;  
doi:10.1093/bioinformatics/bti629

**Archæal genome browser:** <http://archaea.ucsc.edu>

**UCSC bioinformatics degree info:**  
<http://www.bme.ucsc.edu/programs/>

