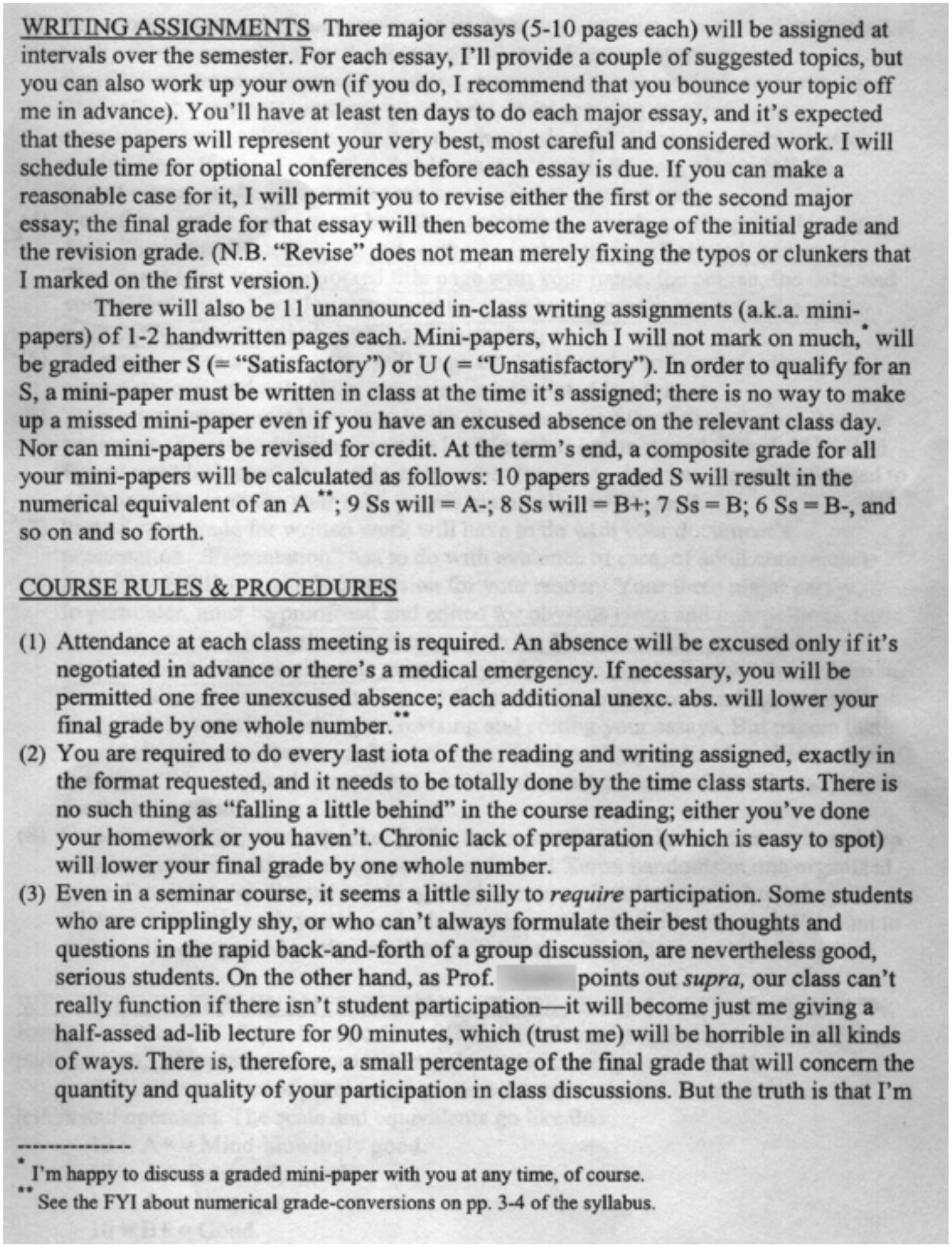
Bacterial Gene Finding

CMSC 423



WRITING ASSIGNMENTS Three major essays (5-10 pages each) will be assigned at intervals over the semester. For each essay, I'll provide a couple of suggested topics, but you can also work up your own (if you do, I recommend that you bounce your topic off me in advance). You'll have at least ten days to do each major essay, and it's expected that these papers will represent your very best, most careful and considered work. I will schedule time for optional conferences before each essay is due. If you can make a reasonable case for it, I will permit you to revise either the first or the second major essay; the final grade for that essay will then become the average of the initial grade and the revision grade. (N.B. "Revise" does not mean merely fixing the typos or clunkers that I marked on the first version.)

There will also be 11 unannounced in-class writing assignments (a.k.a. mini-papers) of 1-2 handwritten pages each. Mini-papers, which I will not mark on much,* will be graded either S (= "Satisfactory") or U (= "Unsatisfactory"). In order to qualify for an S, a mini-paper must be written in class at the time it's assigned; there is no way to make up a missed mini-paper even if you have an excused absence on the relevant class day. Nor can mini-papers be revised for credit. At the term's end, a composite grade for all your mini-papers will be calculated as follows: 10 papers graded S will result in the numerical equivalent of an A **; 9 Ss will = A-; 8 Ss will = B+; 7 Ss = B; 6 Ss = B-, and so on and so forth.

COURSE RULES & PROCEDURES

(1) Attendance at each class meeting is required. An absence will be excused only if it's negotiated in advance or there's a medical emergency. If necessary, you will be permitted one free unexcused absence; each additional unexc. abs. will lower your final grade by one whole number.**

(2) You are required to do every last iota of the reading and writing assigned, exactly in the format requested, and it needs to be totally done by the time class starts. There is no such thing as "falling a little behind" in the course reading; either you've done your homework or you haven't. Chronic lack of preparation (which is easy to spot) will lower your final grade by one whole number.

(3) Even in a seminar course, it seems a little silly to *require* participation. Some students who are cripplingly shy, or who can't always formulate their best thoughts and questions in the rapid back-and-forth of a group discussion, are nevertheless good, serious students. On the other hand, as Prof. [redacted] points out *supra,* our class can't really function if there isn't student participation—it will become just me giving a half-assed ad-lib lecture for 90 minutes, which (trust me) will be horrible in all kinds of ways. There is, therefore, a small percentage of the final grade that will concern the quantity and quality of your participation in class discussions. But the truth is that I'm

---

* I'm happy to discuss a graded mini-paper with you at any time, of course.

** See the FYI about numerical grade-conversions on pp. 3-4 of the syllabus.

Is it O.K. to be a Luddite?

The New York Times Book Review
28 October 1984, pp. 1, 40-41.

As if being 1984 weren't enough, it's also the 25th anniversary this year of C. P. Snow's famous Rede lecture, "The Two Cultures and the Scientific Revolution," notable for its warning that intellectual life in the West was becoming polarized into "literary" and "scientific" factions, each doomed not to understand or appreciate the other. The lecture was originally meant to address such matters as curriculum reform in the age of Sputnik and the role of technology in the development of what would soon be known as the third world. But it was the two-culture formulation that got people's attention. In fact it kicked up an amazing row in its day. To some already simplified points, further reductions were made, provoking certain remarks, name-calling, even intemperate rejoinders, giving the whole affair, though attenuated by the mists of time, a distinctly cranky look.

Today nobody could get away with making such a distinction. Since 1959, we have come to live among flows of data more vast than anything the world has seen. Demystification is the order of our day, all the cats are jumping out of all the bags and even beginning to mingle. We immediately suspect ego insecurity in people who may still try to hide behind the jargon of a specialty or pretend to some data base forever "beyond" the reach of a layman. Anybody with the time, literacy, and access fee can get together with just about any piece of specialized knowledge s/he may need. So, to that extent, the two-cultures quarrel can no longer be sustained. As a visit to any local library or magazine rack will easily confirm, there are now so many more than two cultures that the problem has really become how to find the time to read anything outside one's own specialty.

David Foster Wallace

WRITING ASSIGNMENTS Three major essays (5-10 pages each) will be assigned at intervals over the semester. For each essay, I'll provide a couple of suggested topics, but you can also work up your own (if you do, I recommend that you bounce your topic off me in advance). You'll have at least ten days to do each major essay, and it's expected that these papers will represent your very best, most careful and considered work. I will schedule time for optional conferences before each essay is due. If you can make a reasonable case for it, I will permit you to revise either the first or the second major essay; the final grade for that essay will then become the average of the initial grade and the revision grade. (N.B. "Revise" does not mean merely fixing the typos or clunkers that I marked on the first version.)

There will also be 11 unannounced in-class writing assignments (a.k.a. mini-papers) of 1-2 handwritten pages each. Mini-papers, which I will not mark on much,* will be graded either S (= "Satisfactory") or U (= "Unsatisfactory"). In order to qualify for an S, a mini-paper must be written in class at the time it's assigned; there is no way to make up a missed mini-paper even if you have an excused absence on the relevant class day. Nor can mini-papers be revised for credit. At the term's end, a composite grade for all your mini-papers will be calculated as follows: 10 papers graded S will result in the numerical equivalent of an A **; 9 Ss will = A-; 8 Ss will = B+; 7 Ss = B; 6 Ss = B-, and so on and so forth.

COURSE RULES & PROCEDURES

(1) Attendance at each class meeting is required. An absence will be excused only if it's negotiated in advance or there's a medical emergency. If necessary, you will be permitted one free unexcused absence; each additional unexc. abs. will lower your final grade by one whole number.**

(2) You are required to do every last iota of the reading and writing assigned, exactly in the format requested, and it needs to be totally done by the time class starts. There is no such thing as "falling a little behind" in the course reading; either you've done your homework or you haven't. Chronic lack of preparation (which is easy to spot) will lower your final grade by one whole number.

(3) Even in a seminar course, it seems a little silly to *require* participation. Some students who are cripplingly shy, or who can't always formulate their best thoughts and questions in the rapid back-and-forth of a group discussion, are nevertheless good, serious students. On the other hand, as Prof. [redacted] points out *supra,* our class can't really function if there isn't student participation—it will become just me giving a half-assed ad-lib lecture for 90 minutes, which (trust me) will be horrible in all kinds of ways. There is, therefore, a small percentage of the final grade that will concern the quantity and quality of your participation in class discussions. But the truth is that I'm

----

* I'm happy to discuss a graded mini-paper with you at any time, of course.

** See the FYI about numerical grade-conversions on pp. 3-4 of the syllabus.

Is it O.K. to be a Luddite?

Thomas Pynchon

The New York Times Book Review
28 October 1984, pp. 1, 40-41.

As if being 1984 weren't enough, it's also the 25th anniversary this year of C. P. Snow's famous Rede lecture, "The Two Cultures and the Scientific Revolution," notable for its warning that intellectual life in the West was becoming polarized into "literary" and "scientific" factions, each doomed not to understand or appreciate the other. The lecture was originally meant to address such matters as curriculum reform in the age of Sputnik and the role of technology in the development of what would soon be known as the third world. But it was the two-culture formulation that got people's attention. In fact it kicked up an amazing row in its day. To some already simplified points, further reductions were made, provoking certain remarks, name-calling, even intemperate rejoinders, giving the whole affair, though attenuated by the mists of time, a distinctly cranky look.

Today nobody could get away with making such a distinction. Since 1959, we have come to live among flows of data more vast than anything the world has seen. Demystification is the order of our day, all the cats are jumping out of all the bags and even beginning to mingle. We immediately suspect ego insecurity in people who may still try to hide behind the jargon of a specialty or pretend to some data base forever "beyond" the reach of a layman. Anybody with the time, literacy, and access fee can get together with just about any piece of specialized knowledge s/he may need. So, to that extent, the two-cultures quarrel can no longer be sustained. As a visit to any local library or magazine rack will easily confirm, there are now so many more than two cultures that the problem has really become how to find the time to read anything outside one's own specialty.

NEWS

Girlfriend Stops Reading David Foster Wallace Breakup Letter At Page 20

FEBRUARY 19, 2003 | ISSUE 39•06

BLOOMINGTON, IL—Claire Thompson, author David Foster Wallace's girlfriend of two years, stopped reading his 67-page breakup letter at page 20, she admitted Monday.

Thompson

"It was pretty good, I guess, but I just couldn't get all the way through," said Thompson, 32, who was given the seven-chapter, heavily footnoted "Dear John" missive on Feb. 3. "I always meant to pick it up again, but then I got busy and, oh, I don't know. He's talented, but his letters can sometimes get a little self-indulgent."

Foster, the award-winning author of *The Broom Of The System* and the 1,079-page *Infinite Jest*, met Thompson in March 2001 through mutual friends.

(The Onion)

Mark V Shaney

Developed by Bruce Ellis & Rob Pike in the 1980s

>From mvs Fri Nov 16 17:11 EST 1984 remote from alice
It looks like Reagan is going to say? Ummm... Oh yes, I was looking for. I'm so glad I remembered it. Yeah, what I have wondered if I had committed a crime. Don't eat with your assessment of Reagon and Mondale. Up your nose with a guy from a firm that specifically researches the teen-age market. As a friend of mine would say, "It really doesn't matter"... It looks like Reagan is holding back the arms of the American eating public have changed dramatically, and it got pretty boring after about 300 games.

People, having a much larger number of varieties, and are very different from what one can find in Chinatowns across the country (things like pork buns, steamed dumplings, etc.) They can be cheap, being sold for around 30 to 75 cents apiece (depending on size), are generally not greasy, can be adequately explained by stupidity. Singles have felt insecure since we came down from the Conservative world at large. But Chuqui is the way it happened and the prices are VERY reasonable.

Can anyone think of myself as a third sex. Yes, I am expected to have. People often get used to me knowing these things and then a cover is placed over all of them. Along the side of the $$ are spent by (or at least for) the girls. You can't settle the issue. It seems I've forgotten what it is, but I don't. I know about violence against women, and I really doubt they will ever join together into a large number of jokes. It showed Adam, just after being created. He has a modem and an autodial routine. He calls my number 1440 times a day. So I will conclude by saying that I can well understand that she might soon have the time, it makes sense, again, to get the gist of my argument, I was in that (though it's a Republican administration).
_-_-_-_-Mark

- "I spent an interesting evening recently with a grain of salt."
- "I hope that there are sour apples in every bushel."

(Wikipedia, 2011)

Finding Signals in DNA

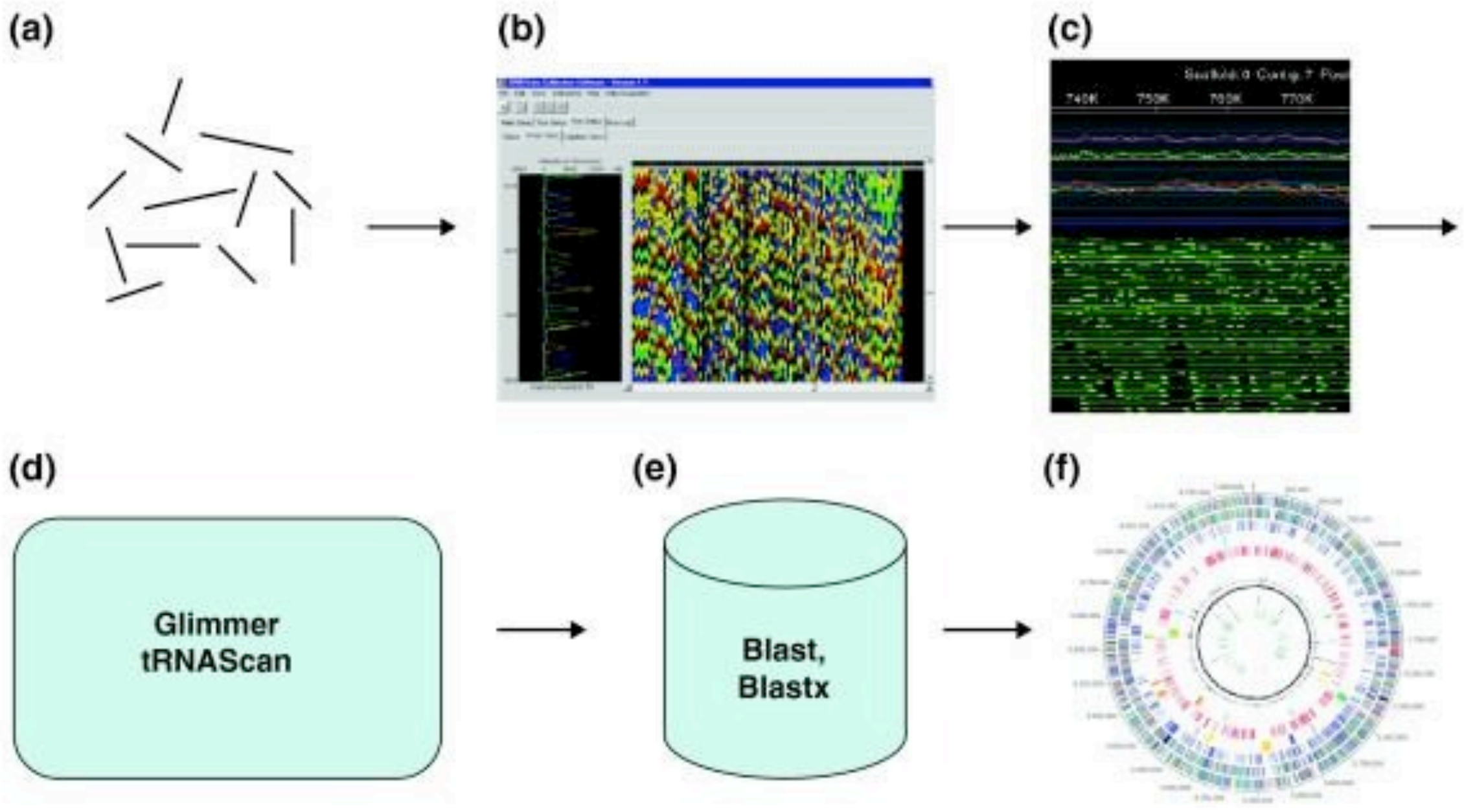
- We just have a long string of A, C, G, Ts. How can we find the "signals" encoded in it?

- Suppose you encountered a language you didn't know. How would you decipher it?

- **Idea #1:** Based on some external information, build a model (like an HMM) for how particular features are encoded.

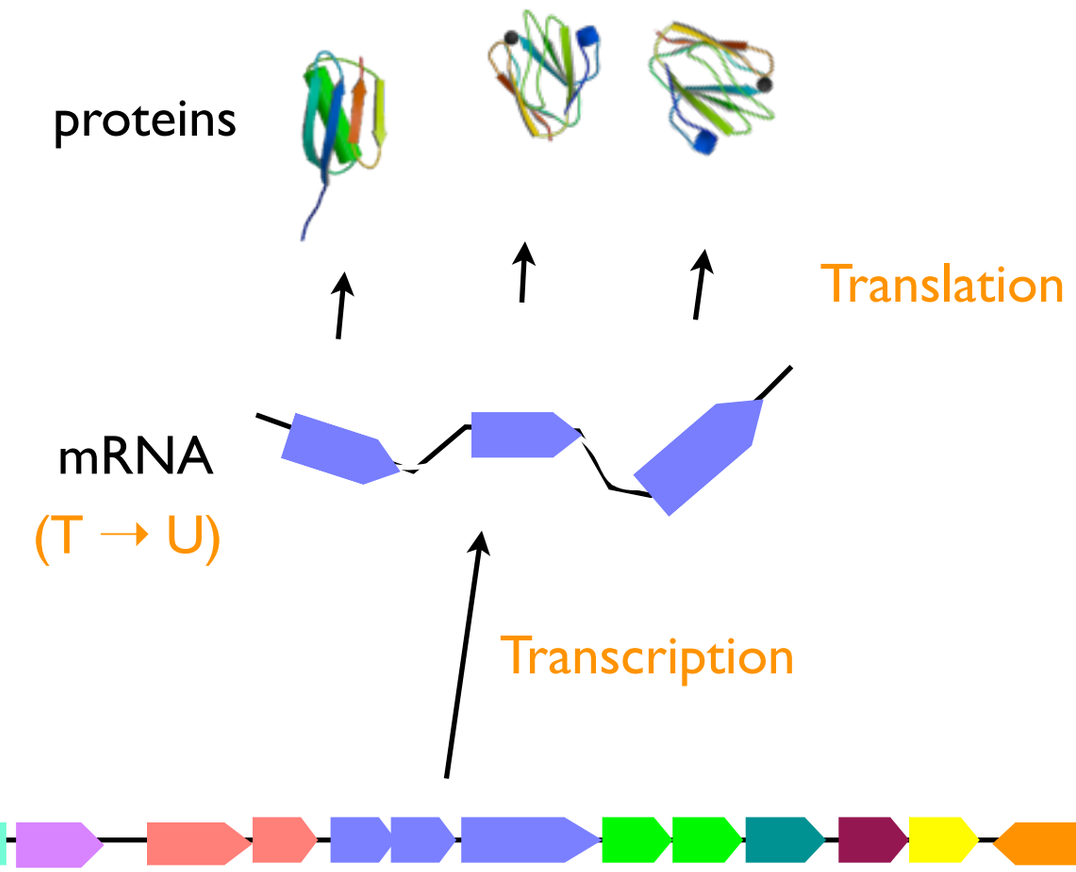
- **Idea #2:** Find patterns that appear more often than you expect by chance. ("the" occurs a lot in English, so it may be a word.)

- Gibbs sampling was an example of how to implement Idea #2.
  We will soon see how to implement idea #1.



Salzberg *Genome Biology* 2007 **8**:102

"Central Dogma" of Biology



DNA =

- double-stranded, linear molecule
- each strand is string over {A,C,G,T}
- strands are complements of each other (A ↔ T; C ↔ G)
- substrings encode for genes  most of which encode for proteins 

1st base	2nd base U		2nd base C		2nd base A		2nd base G	
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine
	UUC	(Phe/F) Phenylalanine	UCC	(Ser/S) Serine	UAC	(Tyr/Y) Tyrosine	UGC	(Cys/C) Cysteine
	UUA	(Leu/L) Leucine	UCA	(Ser/S) Serine	UAA	Ochre *Stop*	UGA	Opal *Stop*
	UUG	(Leu/L) Leucine	UCG	(Ser/S) Serine	UAG	Amber *Stop*	UGG	(Trp/W) Tryptophan
C	CUU	(Leu/L) Leucine	CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine
	CUC	(Leu/L) Leucine	CCC	(Pro/P) Proline	CAC	(His/H) Histidine	CGC	(Arg/R) Arginine
	CUA	(Leu/L) Leucine	CCA	(Pro/P) Proline	CAA	(Gln/Q) Glutamine	CGA	(Arg/R) Arginine
	CUG	(Leu/L) Leucine	CCG	(Pro/P) Proline	CAG	(Gln/Q) Glutamine	CGG	(Arg/R) Arginine
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine
	AUC	(Ile/I) Isoleucine	ACC	(Thr/T) Threonine	AAC	(Asn/N) Asparagine	AGC	(Ser/S) Serine
	AUA	(Ile/I) Isoleucine	ACA	(Thr/T) Threonine	AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
	AUG [A]	(Met/M) Methionine	ACG	(Thr/T) Threonine	AAG	(Lys/K) Lysine	AGG	(Arg/R) Arginine
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine
	GUC	(Val/V) Valine	GCC	(Ala/A) Alanine	GAC	(Asp/D) Aspartic acid	GGC	(Gly/G) Glycine
	GUA	(Val/V) Valine	GCA	(Ala/A) Alanine	GAA	(Glu/E) Glutamic acid	GGA	(Gly/G) Glycine
	GUG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

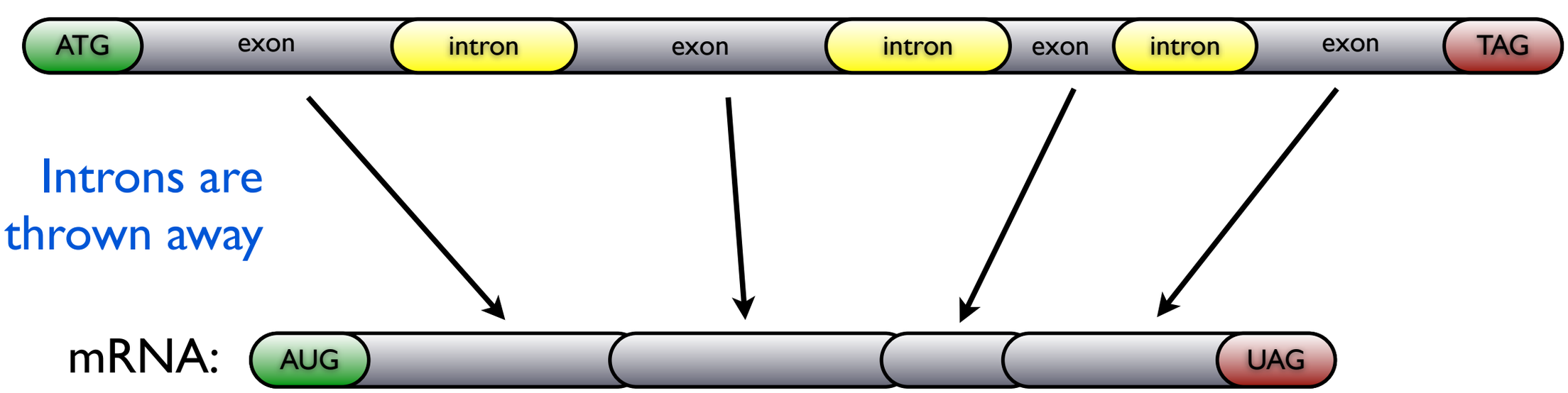
The Genetic Code

- There are 20 different amino acids & 64 different codons.
- Lots of different ways to encode for each amino acid.
- The 3rd base is typically less important for determining the amino acid
- Three different "stop" codons that signal the end of the gene
- Start codons differ depending on the organisms, but AUG is often used.

Eukaryotic Genes & Exon Splicing

Prokaryotic (bacterial) genes look like this:

Eukaryotic genes usually look like this:



This spliced RNA is what is translated into a protein.

The Prokaryotic Gene Finding Problem

- Genes are subsequences of DNA that (generally) tell the cell how to make specific proteins.

- How can we find which subsequences of DNA are genes?

```
Start Codon: ATG
Stop Codons: TGA, TAG, TAA
```

Challenges:

- The start codon can occur in the middle of a gene (where it encodes for the amino acid methionine)
- The stop codon can occur in nonsense DNA between genes.
- The stop codon can occur "out of frame" inside a gene.
- Don't know what "phase" the gene starts in.

A Simple Gene Finder

1. Find all stop codons in genome

2. For each stop codon, find the in-frame start codon farthest upstream of the stop codon, without crossing another in-frame stop codon.

GGC **TAG** **ATG** AGG GCT CTA ACT **ATG** GGC GCG **TAA**

Each substring between the start and stop codons is called an ORF
"open reading frame"

3. Return the "long" ORF as predicted genes.

3 out of the 64 possible codons are stop codons ⇒ in random DNA, every 22nd codon is expected to be a stop.

Gene Finding as a Machine Learning Problem

- Given training examples of some known genes, can we distinguish ORFs that are genes from those that are not?

- **Idea**: can use distribution of codons to find genes.
  - every codon should be about equally likely in non-gene DNA.
  - every organism has a slightly different bias about how often certain codons are preferred.
  - could also use frequencies of longer strings (k-mers).

Bacillus anthracis (anthrax) codon usage

UUU F 0.76	UCU S 0.27	UAU Y 0.77	UGU C 0.73
UUC F 0.24	UCC S 0.08	UAC Y 0.23	UGC C 0.27
UUA L 0.49	UCA S 0.23	UAA * 0.66	UGA * 0.14
UUG L 0.13	UCG S 0.06	UAG * 0.20	UGG W 1.00
CUU L 0.16	CCU P 0.28	CAU H 0.79	CGU R 0.26
CUC L 0.04	CCC P 0.07	CAC H 0.21	CGC R 0.06
CUA L 0.14	CCA P 0.49	CAA Q 0.78	CGA R 0.16
CUG L 0.05	CCG P 0.16	CAG Q 0.22	CGG R 0.05
AUU I 0.57	ACU T 0.36	AAU N 0.76	AGU S 0.28
AUC I 0.15	ACC T 0.08	AAC N 0.24	AGC S 0.08
AUA I 0.28	ACA T 0.42	AAA K 0.74	AGA R 0.36
AUG M 1.00	ACG T 0.15	AAG K 0.26	AGG R 0.11
GUU V 0.32	GCU A 0.34	GAU D 0.81	GGU G 0.30
GUC V 0.07	GCC A 0.07	GAC D 0.19	GGC G 0.09
GUA V 0.43	GCA A 0.44	GAA E 0.75	GGA G 0.41
GUG V 0.18	GCG A 0.15	GAG E 0.25	GGG G 0.20

An Improved Simple Gene Finder

- Score each ORF using the product of the probability of each codon:

$$GFScore(g) = Pr(codon_1) x Pr(codon_2) x Pr(codon_3) x ... x Pr(codon_n)$$

But: as genes get longer, GFScore(g) will decrease.

So: we should calculate GFScore(g[i...i+k]) for some window size k.

The final GFSCORE(g) is the average of the Scores of the windows in it.

Recap

- Simple gene finding approaches use codon bias and long ORFs to identify genes.

- Many top gene finding programs for Eukaryotes are based on generalizations of Hidden Markov Models because multiple types of signals (many "authors") are present in a gene (intron, exon, etc.)

- Basic HMMs must be generalized to emit variable sized strings.