I519: Introduction to Bioinformatics, Fall, 2015

Aug 28th, 2015

Homework 1 (Due via canvas by **11:59pm on Thursday, Sep 10th, 2015**)

Guidance on programming assignments:

1) Only turn in your OWN codes.
2) Use comments in your programs.
3) If the program names are given in the assignment, stick with the given names. Otherwise, you are responsible for giving your program a nice name.
4) Include README file for each programming assignment. For example, if your program is test.py, you need to include a test.readme file. The README file is not supposed to be lengthy, but should contain concrete and enough information: function of your program, its input/output, and sample usage.
5) Do **NOT** use biopython for the assignments**.**
6) Use burrow.soic.indiana.edu to write/test your codes, as your codes are likely to be examined on this server for grading.
7) You can find all input files (and this instruction) in burrow.soic.indiana.edu under /u/yye/I519/HW1/. Don't copy large genome files to your own folder. Make a symbolic link (using ln -s) instead.

1. Write a simple python code for guessing the type (DNA, RNA or Protein) of input sequences in a given FASTA file (40pts)

Write a program **GuessType.py** that can tell the type of biological sequence(s) (DNA, RNA or Protein) given in a FASTA file. It also reports the frequencies of the different letters found in the sequence(s). The input of your program is a FASTA file with a single or multiple sequences, and your program outputs the type of the sequence(s), and the frequencies of the different letters. Make sure that you don't hard code the name of the input sequence file; instead, the name should be given as a command-line option. You can use **seq1.fa**, **seq2.fa** and **seq3.fa** for testing your code.

2. Data analysis using R (40pts)

As I mentioned in the class that due to the advances of sequencing technology, we now may have access to thousands of genomes for a single species. For example, there are more than four thousands of *Staphylococcus aureus* genomes available at the NCBI website. Only a small fraction of these genomes are complete, while the remaining ones are draft (not completely assembled). Draft genomes are good for many analyses. The first analysis I have done is to predict the genes in all the genomes, and count the total

number of genes I found in each of the genomes. I have prepared two files with information of gene numbers (each line shows a strain and the total number of genes found in the strain): one for complete genomes (**genesize-complete.txt**), and the other one for draft genomes (**and genesize-draft.txt)**. I hope you will help me make sense of these data. First, I'd like to know the average number of genes in each genome. Second, I want to know if draft genomes and complete genomes have similar numbers of genes. If your answer to the second question is "no", you need to provide me some statistical evidence to convince me that they are really different. Finally, I am very picky—I'd like to hear some explanations to the difference if you find any. Use R for the analyses (and plotting) for this assignment.

3. Paper reading (20 pts).

For this assignment, you are going to read an article recently published in Time—so hopefully it will be fun to read. Sorry I don't have the electronic copy of the article. I scanned it, so the resolution is not super as you can see below. Write a summary of this article, in anyway you like. [But we will have rules for paper readings in the future].

A strange new way to solve crimes

The bacteria growing in and on the human body is so unique—and so revealing—that scientists believe germs will soon help catch bad guys

By Mandy Oaklander

1 *South Florida detectives drive CSI vans to a mock crime scene where a novel new tool in forensics will be tested*

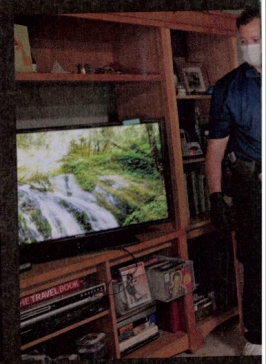
Staging a burglary in the name of science, detectives break into a house through a front window

2

Scientists later swab the window to collect and analyze invisible bacteria left behind by the "burglars"

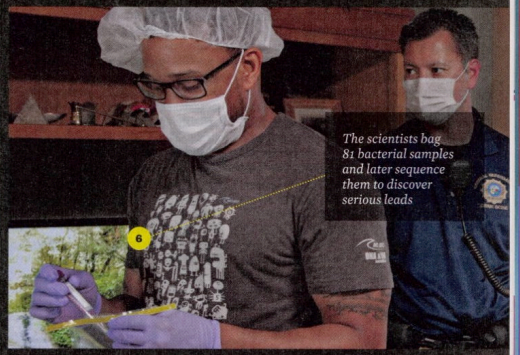
3

4 *Switches and other surfaces that were touched in the mock crime also get swabbed*



Uniformed members of the crime-scene unit wear sterile gear to prevent contaminating the scene

5



The scientists bag 81 bacterial samples and later sequence them to discover serious leads

6

PHOTOGRAPHS BY CHRISTOPHER MORRIS FOR TIME

FEW THIEVES ARE silly enough to burgle a house in broad Florida daylight—let alone a house where three crime-scene-investigation police vans are already parked out front. This is obviously Bill Stewart's first time as a criminal. "Can I crawl through your window?" he asks the homeowner. "If it doesn't break anything?"

Stewart is a detective sergeant in the Fort Lauderdale police department, and though he might be a failure as a robber, he knows exactly how real ones operate—and that's why he's here. In his 26 years on the force, he's searched hundreds of windowsills, garden tools and hastily discarded gloves for clues about whodunit. Now, by playacting the role of a typical burglar, he's participating in an unusual scientific study that could ultimately change how crimes get solved.

But first he and his partner need to muck up the place, which belongs to a scientist involved in the study. They squeeze through a small square window—a popular point of entry for burglars—and once inside, split up and go looking for valuables. They switch on the lights and rummage through drawers. One raids the fridge and drinks half a Diet Coke. They grab a pillowcase to stuff their loot in—something robbers often do, says Stewart. Then they sit on the couch next to the family cat, Sammie, and slip an iPad and a laptop into the pillowcase before yanking the TV's cords from the wall to cart it away.

After they leave, scientists in sterile gear file in. Led by Jarrad Hampton-Marcell, a research coordinator at the Argonne National Laboratory, which works for the U.S. Department of Energy, a team goes room by room collecting cotton-tipped swabs of what the robbers left behind.

So far, this crime-scene activity looks as routine as the pacing of a television police procedural. But there's a twist. They won't be gathering and analyzing DNA or fingerprints. They'll be analyzing bacterial cells left behind by the robbers.

Think of it as CSI: *E. coli*. New science is finding that each one of us brings with us (and can't help but leave behind) a unique bacterial signature everywhere we go—a germy John Hancock. As you move through a scene and shed your microbes, the space starts to reflect your bacterial signature, potentially tying you to it and giving away a lot about you in the process.

To test how much bacteria gets left behind and what it can reveal about identity, scientists will compare the swabs collected from the robbers and see if they can differentiate them from those of the homeowners and their cat, whose paw the scientists also swabbed. They'll also try to see if they can tease out the signatures from the samples from the scene. If they can, it will provide early proof that an outsider's bacteria is distinct enough from the homeowners' to confirm that a stranger was in the house.

If this holds true, and evidence suggests it might, it would mean crime scenes are riddled with valuable clues that are currently left untested. Crime experts agree that the field of forensics needs cheaper, faster ways to gather investigative leads like these. Trace evidence—the kind found through hairs, fibers or paint—typically requires chemical analysis, which can be expensive and inaccurate if there's not enough of it to analyze. Thanks to advances in science, however, bacterial evidence can be sequenced affordably, quickly and with startling accuracy. That's why forensics experts are saying it's the leading contender for next-generation investigations.

"The criminal-justice system is always looking for one thing: they're looking for probable cause, any kind of thing that can give them information about a possible suspect," says George Duncan, DNA unit manager at the Broward County sheriff's office crime lab, who has worked in forensics for 43 years. "The crime-scene people, they think bacterial forensics is just as exciting as hell."

YOUR BACTERIAL MAKEUP, called the microbiome, can give away a lot about you. So far, most research has focused on the densest site for bacteria, the gut, which houses roughly four pounds of bugs. But bacteria isn't just in us; we're covered in the stuff too.

By age 3, everyone, even identical twins, has a unique coat of it that changes somewhat but remains largely consistent at its core and over time. In scientific studies, researchers have successfully matched smartphones and keyboards to the people who used them by analyzing their microbial signatures. And in a study published last year in the journal *Science*, researchers followed seven families for six weeks and were able to match them to their homes through bacteria alone. The more intimate you are with someone, the study found, the more microbes you share—though your makeup is still distinct from theirs.

"There's a continuum between you and your world, not a brick wall that ends at your skin," says Jack Gilbert, a microbial ecologist at Argonne and principal investigator of that study. He and his team discovered that even when a family moved, it took only hours for the new house to look nearly bacterially identical to the old one.

Scientists' ability to track bacteria left behind by people is where the forensic potential lies. Gilbert, who's studied microbes for 16 years, thinks bacterial forensics will be the next great contribution to crime fighting. "We're ramping up to be able to leverage signature profiles in a really robust way," he says. "It's what people did for fingerprints years ago."

Court challenges will follow the scientific ones—it took a decade for DNA to be a courtroom staple—but here's how it could play out: in a case like murder, the prime suspects are usually people closest to the victim. So if a wife is killed, says Stewart Mosher, a sergeant with the Broward County sheriff's office crime-scene unit, "the first person you've got to look at is the husband." Consider, however, that the husband says he was out of town when it happened. Bacterial signatures last 48 to 72 hours once a person has left. "So if his bacterial profile is absent from the house, and that matches his sworn statement, which we would have to substantiate, it's going to be extremely difficult to be able to say he had anything to do with it. That clue alone could be huge."

It's a brand-new area of physical evidence, says David Carter, a forensic

Even Sammie the cat got swabbed. Research suggests that a pet can alter the bacterial profile of a home

CHRISTOPHER MORRIS—VII FOR TIME

specialist who assists the Honolulu police. "We've lacked science and technology to analyze microbial communities," he says. But with fast new ways to sequence microbes without having to grow them in a lab, "now we can get a level of resolution that we never had before."

Of course, there's a chasm between the potential of bacterial forensics and its widespread adoption. Some legal experts cast doubt on how reliable the technique is—and how useful, given that DNA would likely be wherever bacteria is present. And what does it mean if the signatures are close, but not identical? These are some of the questions that need answering before it's admissible in court. "I don't see this, so far, as revolutionizing forensic science," says David H. Kaye, a law professor at Pennsylvania State University and forensic-science expert. On top of all that, there are also some hairy ethical questions to be grappled with first.

DEPENDING ON HOW it's sequenced, a microbial sample can reveal private information about its host, from what diseases they might have to what kind of work they do to their ethnicity. That information, some caution, is far too sensitive to put into a database. But Gilbert doesn't see how it's ethically different from collecting genetic information left at a crime scene. "It may come to a point where, if you perform a criminal act, you have your microbiome collected and databased," he says. "We're a long way off that, but it's something I would like to work toward."

In a study, scientists have successfully matched smartphones to their owners simply by analyzing the bacteria present

In the meantime, thousands of volunteers are willingly sharing their bacterial signatures with researchers. Large-scale databases are sequencing their microbiomes, and scientists are finding valuable correlations by comparing the bacteria of one person against a database of others.

For instance, after analyzing the bacteria collected from the faux burglary, Hampton-Marcell found that the robbers' bacteria were indicative of two quirky factors: regular drinking and migraines. The owners of the house, another comparison revealed, were omnivores and popped vitamin B and calcium. (Turns out Stewart does get migraines, and the homeowners do eat everything—including vitamins.) So, in addition to proving a stranger has been in a home, scientists theorize that bacteria could also tell investigators more about what kind of a person the suspect is.

Cops won't be swabbing for bacteria tomorrow. But, says Kaye, "I can imagine some cases where this starts to be used for investigative purposes in five to ten years."