Comparing Bioinformatics Software Development by Computer Scientists and Biologists

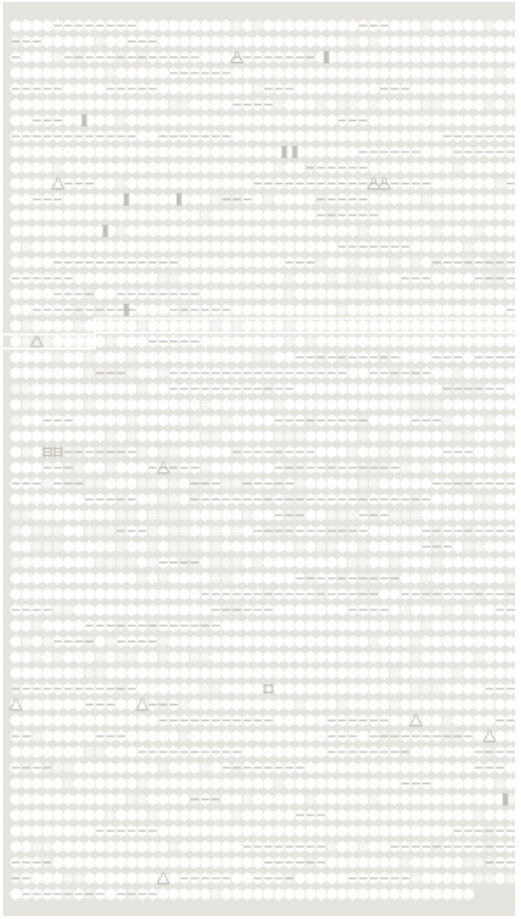
An Exploratory Study

Parmit K. Chilana, Andrew J. Ko
The Information School
University of Washington

Carole L. Palmer
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

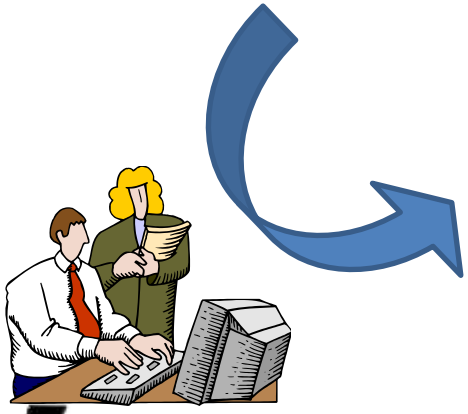
AATTCATTTTTAATCCTTTAATAG
TCCACAGTAATATTGTCCTAAAGA
GGGTACATTGGATTTTAATTTTGC
TTTCAATATGACGGCTGTCAATGT
TGCCCTGATTCGTGATACCAAGTG
GCTGACTTTAGAAGTCTGTAGAGA
ATTTCAGAGAGGAACTTGCTCTCG
AGCTGATGCAGATTGCAAGTTTGC
CCATCCACCAAGAGTTTGCCATGT
GGAAAATGGTCGTGTGGTGGCCTG
TTTTGATTCTCTAAAGGGTCGGTG
TACCCGAGAGAACTGCAAGTACCT
TCACCCTCCTCCACACTTAAAAAC
GCAGCTGGAGATTAATGGGCGGAA
CAATCTGATTCAACAGAAGACTGC
CGCAGCCATGTTCGCCCAGCAGAT
GCAGCTTATGCTCCAAAACGCTCA
AATGTCATCACTTGGTTCTTTTCC
TATGACTCCATCAATTCCAGCTAA
TCCTCCCATGGCTTTCAATCCTTA
CATACCACATCCTGGGATGGGCCT
CGTTCCTGCAGAACTTGTACCAAA

```
print "This script will convert
your DNA sequence to RNA
Sequence\n\n";
print "ENTER THE FILENAME OF
THE DNA SEQUENCE:= ";
$dnafilename = <STDIN>;
chomp $dnafilename;
unless ( open(DNAFILE,
$dnafilename) ) {
    print "Cannot open file
\"$dnafilename\"\n\n";
    goto h;
}
@DNA = <DNAFILE>;
close DNAFILE;
$DNA = join( '', @DNA);
print "The original DNA
Sequence :=\n\n";
$DNA =~ s/\s//g;
print "$DNA\n\n";
$RNA = $DNA;
$RNA =~ s/T/U/g;
$RNA =~ s/t/u/g;
print "Transcribing DNA TO RNA
:=\n\n";
print "$RNA\n";
<STDIN>;
```

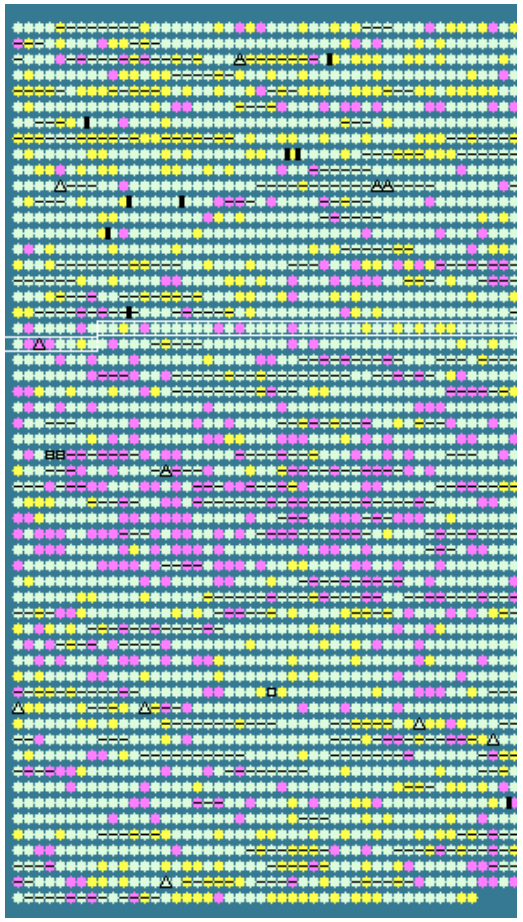


Bioinformatics

```
AATTCATTTTTAATCCTTTAATAG
TCCACAGTAATATTGTCCTAAAGA
GGGTACATTGGATTTTAATTTTGC
TTTCAATATGACGGCTGTCAATGT
TGCCCTGATTCGTGATACCAAGTG
GCTGACTTTAGAAGTCTGTAGAGA
ATTTCAGAGAGGAACTTGCTCTCG
AGCTGATGCAGATTGCAAGTTTGC
CCATCCACCAAGAGTTTGCCATGT
GGAAAATGGTCGTGTGGTGGCCTG
TTTTGATTCTCTAAAGGGTCGGTG
TACCCGAGAGAACTGCAAGTACCT
TCACCCTCCTCCACACTTAAAAAC
GCAGCTGGAGATTAATGGGCGGAA
CAATCTGATTCAACAGAAGACTGC
CGCAGCCATGTTCGCCCAGCAGAT
GCAGCTTATGCTCCAAAACGCTCA
AATGTCATCACTTGGTTCTTTTCC
TATGACTCCATCAATTCCAGCTAA
TCCTCCCATGGCTTTCAATCCTTA
CATACCACATCCTGGGATGGGCCT
CGTTCCTGCAGAACTTGTACCAAA
```



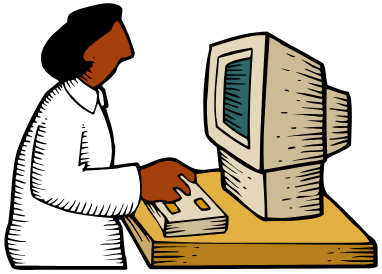
```
print "This script will convert
your DNA sequence to RNA
Sequence\n\n";
print "ENTER THE FILENAME OF
THE DNA SEQUENCE:= ";
$dnafilename = <STDIN>;
chomp $dnafilename;
unless ( open(DNAFILE,
$dnafilename) ) {
    print "Cannot open file
\"$dnafilename\"\n\n";
    goto h;
}
@DNA = <DNAFILE>;
close DNAFILE;
$DNA = join( '', @DNA);
print "The original DNA
Sequence :=\n\n";
$DNA =~ s/\s//g;
print "$DNA\n\n";
$RNA = $DNA;
$RNA =~ s/T/U/g;
$RNA =~ s/t/u/g;
print "Transcribing DNA TO RNA
:=\n\n";
print "$RNA\n";
<STDIN>;
```



information needs and task analysis

Computer Scientists

Biologists

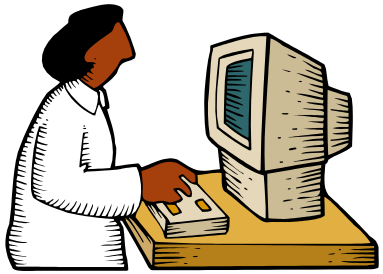


surveys of practice

end-user programming

software development

Biologists



Computer Scientists

software development

Biologists

Computer Scientists

information-seeking activities

Research Questions

- What type of information do bioinformatics software developers seek?
- Where do they obtain this information?
- Are there differences between the information-seeking activities of biologists vs. computer scientists?

The Study

- Semi-structured interviews
- 8 participants, 4 different labs
- Analysis: grounded, inductive approach

CS Researchers

probability theory
machine learning
statistical analysis

MBB Researchers

large genomic data analysis
visualization of patterns

CS Practitioners

Processing raw data
Supporting data transformations
Database maintenance

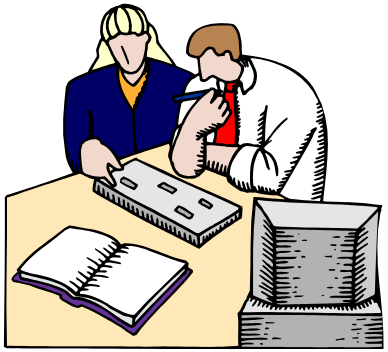
MBB Practitioners

Processing raw data
Supporting data transformations
Database maintenance

Stages of development

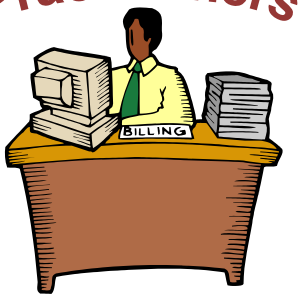
Understanding the Problem → **Translating the Problem into Code** → **Interpreting Results**

```
print "This script will convert
your DNA sequence to RNA
Sequence\n\n";
print "ENTER THE FILENAME OF
THE DNA SEQUENCE:= ";
$dnafilename = <STDIN>;
chomp $dnafilename;
unless ( open(DNAFILE,
$dnafilename) ) {
    print "Cannot open file
\"$dnafilename\"\n\n";
    goto h;
}
@DNA = <DNAFILE>;
close DNAFILE;
$DNA = join( '', @DNA);
print "The original DNA
Sequence :=\n\n";
$DNA =~ s/\s//g;
print "$DNA\n\n";
```



Working with structured vs. ill-structured problems

Practitioners



- practitioners appeared to work with well-structured problems
- interested in "instrumental" information – how to implement a particular solution

Researchers

- researchers in both categories frequently faced ill-structured problems – exploratory, complex
- goal was to break down the problem into smaller, structured problems

Understanding contributing factors and interrelationships

Biologists

- understood the biological relevance, but sought information to understand the technical implementation
- Ie. which data structure to use

Computer Scientists

- once they had developed their algorithms and code, they sought information to ensure the biological relevance
- i.e., cut-off values in function parameters

Dealing with familiar vs. new problems

Biologists

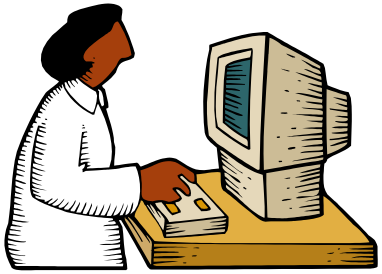
- favor familiar programming languages and platforms, but comfortable with tackling new biological problems
- looking up information about a new technology was not sufficient – needed coaching

Computer Scientists

- able to apply programming and debugging skills in different situations, but struggled with new biological contexts
- reading up on the relevant biological details was not enough – needed to talk to biologists

Making sense of output

Biologists



- could often determine if an output was valid based on inspection
- sought information to understand more complex problems or to revise their programs to do more sophisticated analyses

Computer Scientists

- could devise a technical solution or an algorithm valid from a CS perspective, but output may or may not be valid biologically

Limitations

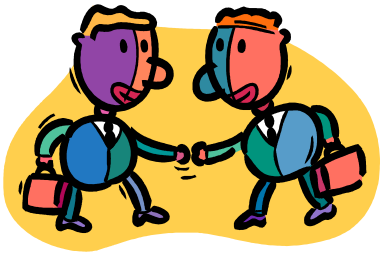
- Small sample size
- Only perspective from North America, academic research labs
- Inherent limitation in the use of semi-structured interview techniques

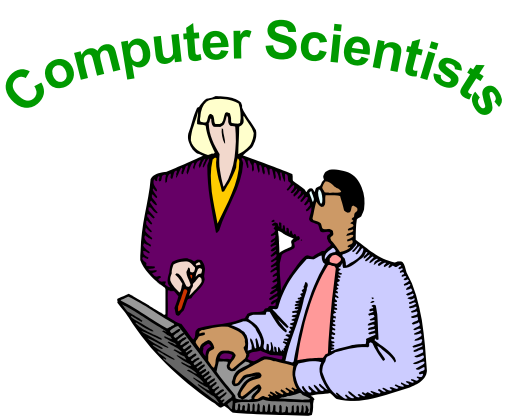
Take-away points

- Biologists want to have control over their analyses but would rather not write code
- Developers need more efficient ways of learning about the domain they are working in
- Clearly a preference for **informal exchange** of information whether it is to solve a technical problem or to learn about the underlying biology

Software development in bioinformatics depends on **collaboration**

Not possible to do this type of work without mutual dependency

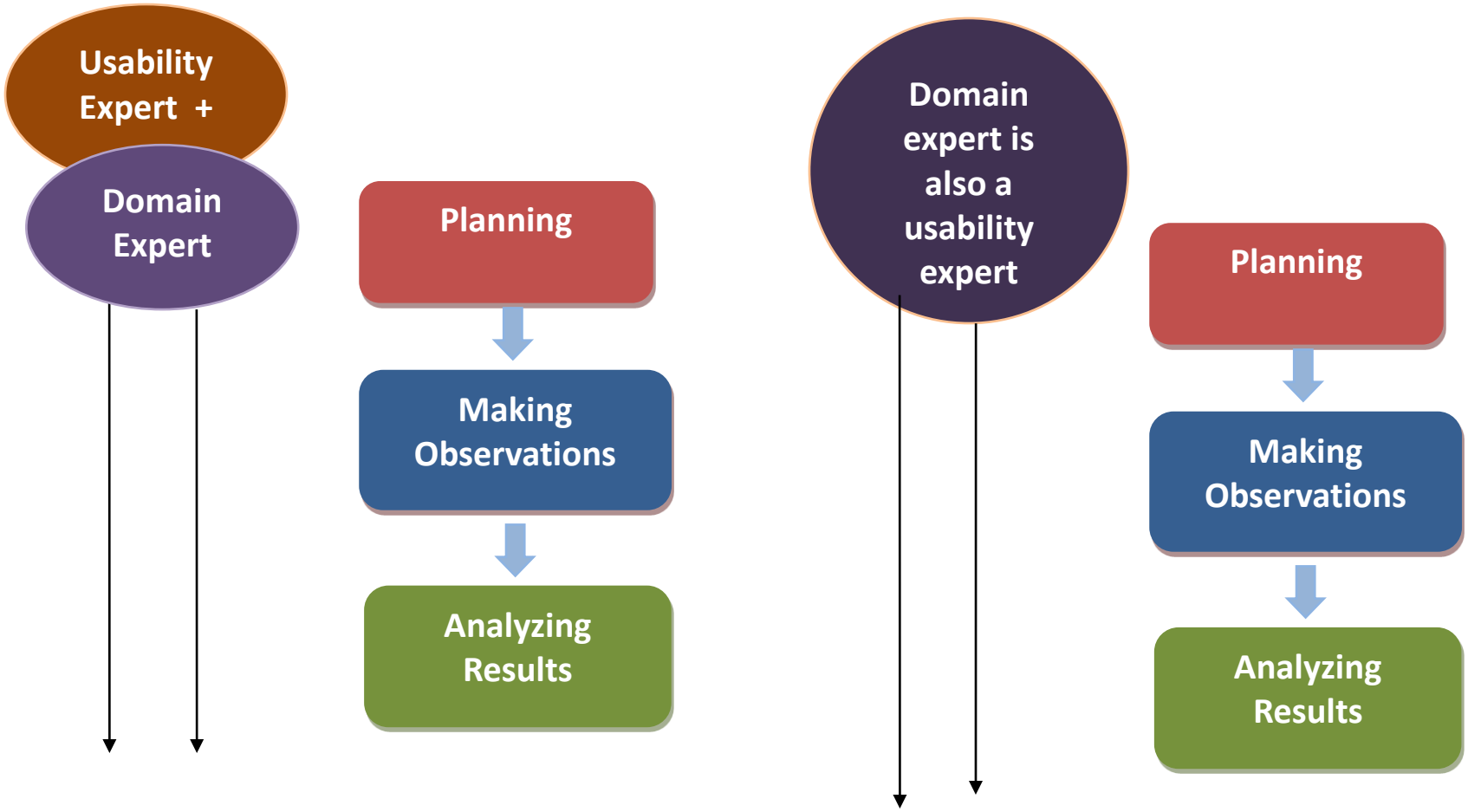




domain expertise and user research

- Another study looking at how user researchers work in unfamiliar domains
- Conducted 21 interviews with in-house usability experts, consultants, and managers
- Variety of domains: medical imaging, financial analysis, software development, network security, measurement instruments
- Challenges in devising and executing usability tests, and analyzing results

domain expertise and user research



Thank-you.

Questions?

Parmit K. Chilana, Carole L. Palmer, Andrew J. Ko
University of Washington
University of Illinois at Urbana-Champaign