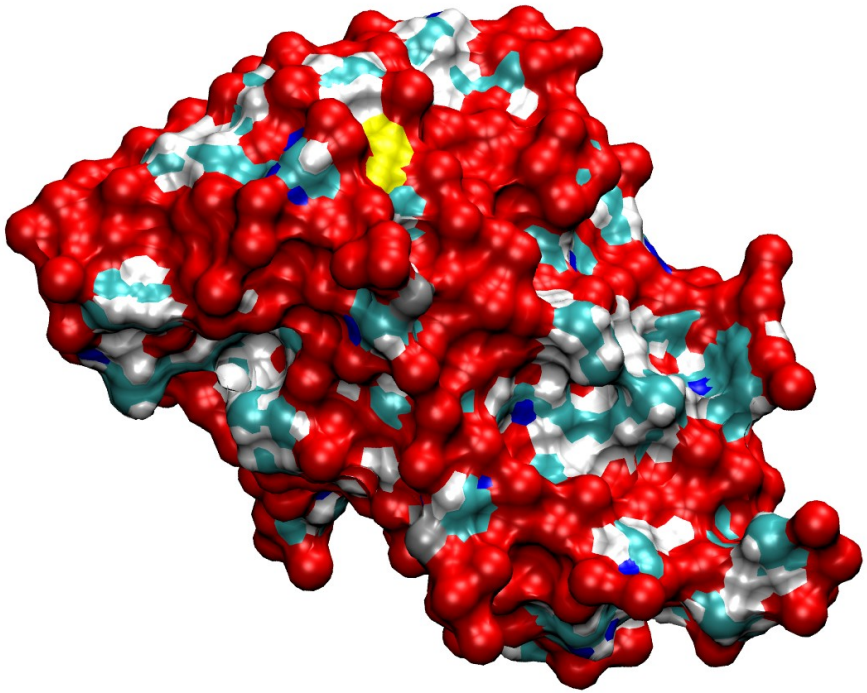
Information Theory and Sequence Logos

Yosi Shibberu

Rose-Hulman Institute of Technology
2014

Insulin



ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Insulin Sequence

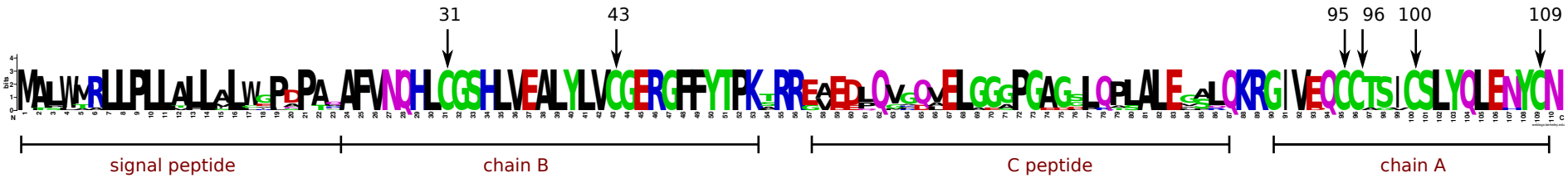
```
MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY    50
TPKTRREAED LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC   100
SLYQLENYCN                                               110
```

Insulin Sequence for 11 Species

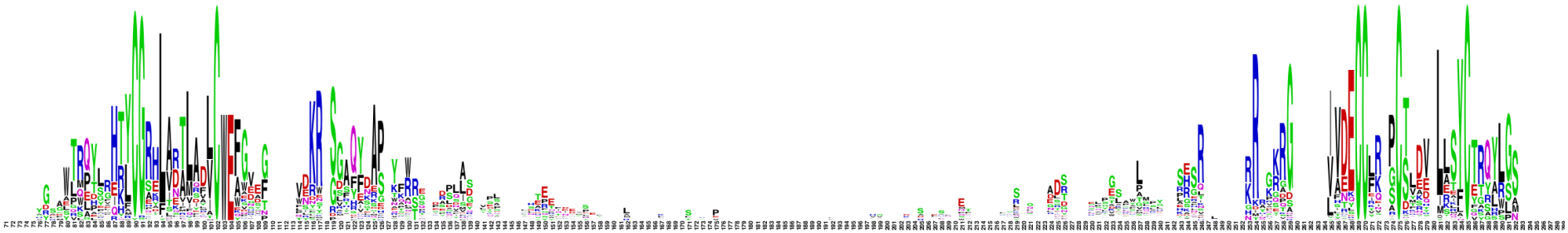
```
dog        P01321 INS_CANFA  1   MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVED   60
hamster    P01313 INS_CRILO  1   MTLWMRLLPLLTLLVLWEPNPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRRGVED   60
cat        P06306 INS_FELCA  1   MAPWTRLLPLLALLSLWIPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAED   60
gorilla    Q6YK33 INS_GORGO  1   MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
human      P01308 INS_HUMAN  1   MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
monkey     P30406 INS_MACFA  1   MALWMRLLPLLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
chimpanzee P30410 INS_PANTR  1   MALWMRLLPLLVLLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
orangutan  Q8HXV2 INS_PONPY  1   MALWMRLLPLLALLALWGPDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
rat        Q62587 INS_PSAOB  1   MALWMRLLPLLAFLILWEPSPAHAFVNQHLCGSHLVEALYLVCGERGFFYTPKFRRGVDD   60
rabbit     P01311 INS_RABIT  1   MASLAALLPLLALLVLCRLDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRREVEE   60
squirrel   Q91XI3 INS_SPETR  1   MALWTRLLPLLALLALLGPDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRREVEE   60
                                 *:    *****.:* *    *: ****************************** ** .::

dog        P01321 INS_CANFA 61   LQVRDVELAGAPGEGGLQPLALEGALQKRGIVEQCCTSICSLYQLENYCN  110
hamster    P01313 INS_CRILO 61   PQVAQLELGGGPGADDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN  110
cat        P06306 INS_FELCA 61   LQGKDAELGEAPGAGGLQPSALEAPLQKRGIVEQCCASVCSLYQLEHYCN  110
gorilla    Q6YK33 INS_GORGO 61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
human      P01308 INS_HUMAN 61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
monkey     P30406 INS_MACFA 61   PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
chimpanzee P30410 INS_PANTR 61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
orangutan  Q8HXV2 INS_PONPY 61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
rat        Q62587 INS_PSAOB 61   PQMPQLELGGSPGAGDLRALALEVARQKRGIVEQCCTGICSLYQLENYCN  110
rabbit     P01311 INS_RABIT 61   LQVGQAELGGGPGAGGLQPSALELALQKRGIVEQCCTSICSLYQLENYCN  110
squirrel   Q91XI3 INS_SPETR 61   QQGGQVELGGGPGAGLPQPLALEMALQKRGIVEQCCTSICSLYQLENYCN  110
                                  *  : **. .**    :  ***   ******:***:.:*******.***
```

Insulin Sequence Logo for 11 Species

Insulin Sequence Logo for 11 Species



Insulin Sequence Logo for 11 Species

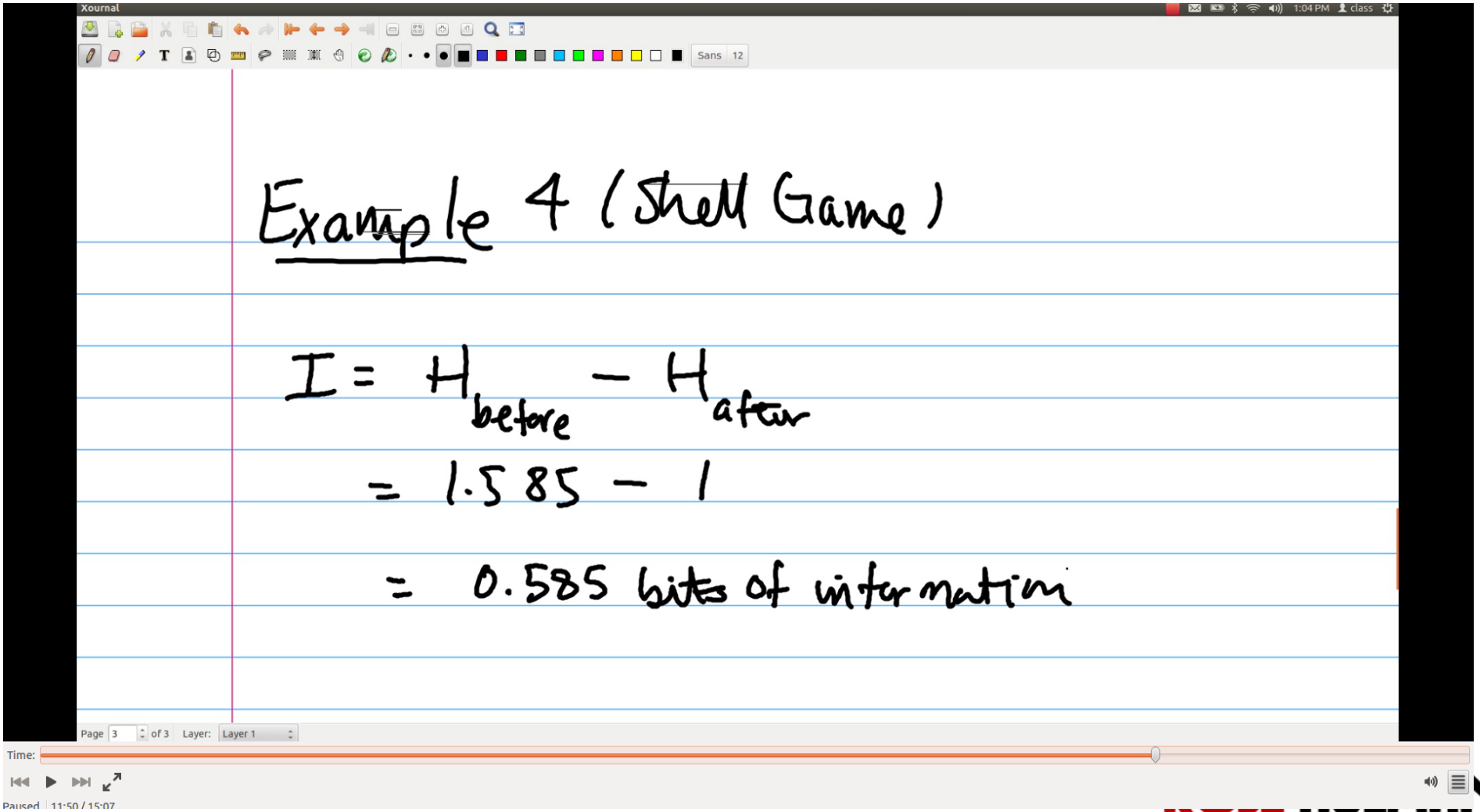


ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Module Materials

- Videos (7 approx 10 min each)
- Handouts (with space for solutions to examples)
- Lessons (3 with solutions for instructors)

Screenshot of a Video



Handouts

4 Example (Shell Game)
Repeat the shell game, but this time assume I choose a cup at random and reveal the cups *does not* contain the token. How much information have I given you?

Solution:

Lessons

Lesson 3 (Password Entropy) The entropy of a password is a measure of how difficult it is to guess the password.

(a) Are passwords with high entropy preferable to passwords with low entropy? Explain.
(b) Compute the entropy of the password `deoqkelrti`. State any assumption.
(c) Compute the entropy of the password `1Ar8s`. State any assumption.
(d) Which password, `deoqkelrti` or `1Ar8ns`, is more secure? Explain.
(e) Is the password, `password` very secure? Explain.

(a) Yes, High entropy passwords have a higher level of uncertainty than low entropy passwords making high entropy passwords harder to guess.

(b) Assuming we limit our choices to the 26 lower case letters since we use 10 letters

$H(X) = -\sum p \log_2 p = -10 \log_2 \frac{1}{26} = 47.0$ bits

(c) Now we can choose from 26 lower, 26 upper letters and 10 numbers we have for passwords of length 5:

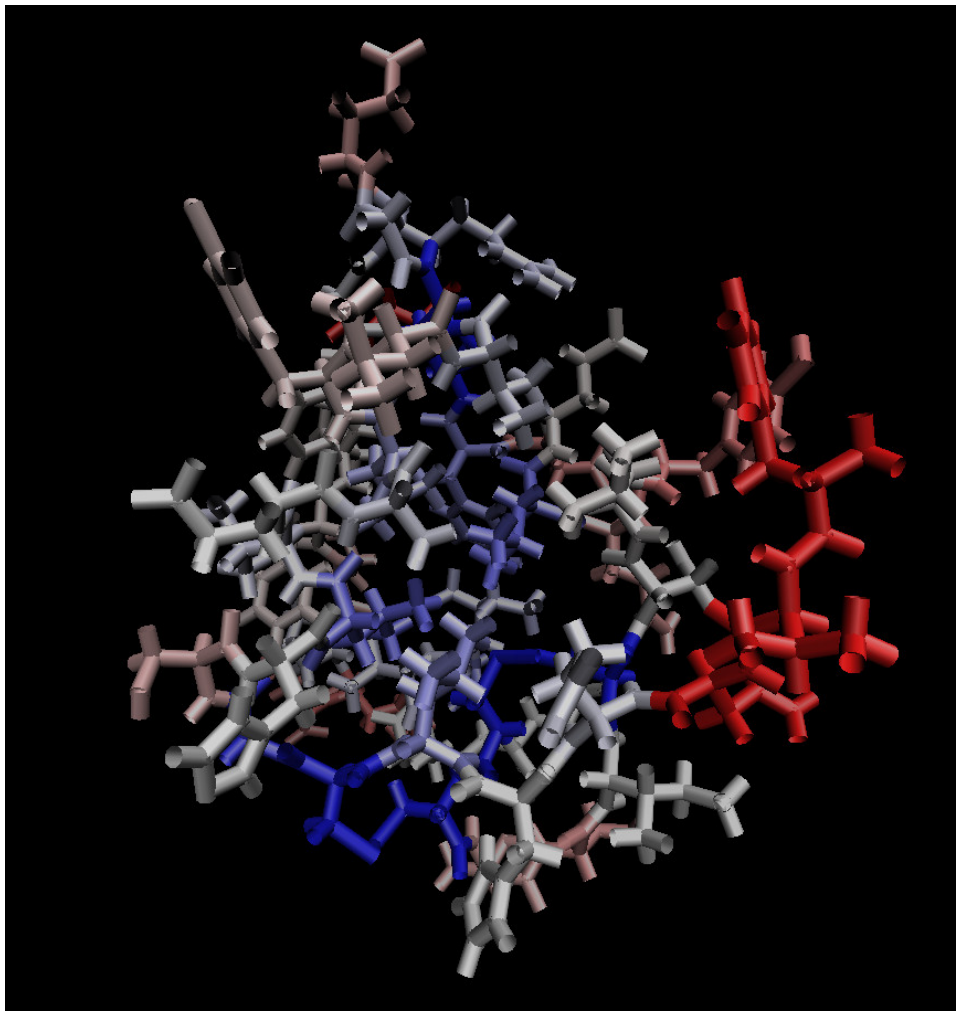
$H(X) = -\sum p \log_2 p = -5 \log_2 \frac{1}{26+26+10} = 29.8$ bits

(d) Since $47.0 > 29.8$ the password deoqkelrti is more secure than 1Ar8ns.

(e) No! A better way for computing the entropy of such passwords is to tally the frequency of commonly used "bad" passwords and use this tally to compute the probabilities needed to determine the entropy.

Student Project (Steven Haussmann)

Insulin Structure Color Coded by Information Content



ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Acknowledgements

- Mark Brandt, Chemistry & Biochemistry Dept.
- Nathanael Moore, CS Senior
- Andrew Piliser, Math Senior
- Rain Dartt, Computer Engineering Junior
- Steven Gregory, Chemistry Freshman

ROSE-HULMAN INSTITUTE OF TECHNOLOGY