Information Transfer in Biological Systems

W. Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907

July 1, 2007

AofA and IT logos

*Joint work with A. Grama, P. Jacquet, M. Koyoturk, M. Regnier, and G. Seroussi.
1. What is Information?
2. Beyond Shannon Information
3. Information Transfer: Darwin Channel
   • Model: Deletion and Constrained Channels
   • Capacity of the Noisy Constrained Channel
4. Information Discovery in Massive Data: Pattern Matching
   • Classification of Pattern Matching Problems
   • Statistical Significance
   • Finding Weak Signals in Biological Data
5. Information in Structures: Network Motifs
   • Biological Networks
   • Finding Biologically Significant Structures
What is Information?

**C. F. Von Weizsäcker:**

“*Information* is only that which *produces information*” (relativity).

“*Information* is only that which *is understood*” (rationality)

“*Information* has no absolute meaning.

**C. Shannon:**

“These *semantic* aspects of communication are *irrelevant* . . .”

**F. Brooks, jr.** (JACM, 50, 2003, “*Three Great Challenges for . . . CS*”):

“*Shannon* and Weaver performed an inestimable service by giving us a definition of *Information* and a metric for for *Information* as *communicated* from place to place.

We have *no theory* however that gives us a metric for the *Information* embodied in *structure* . . .

this is the most *fundamental gap* in the theoretical underpinning of *Information* and computer science. . . . A young information theory scholar willing to spend years on a *deeply fundamental problem* need look no further.”
Some Definitions

Information has flavor of:

- **relativity** (depends on the activity undertaken),
- **rationality** (depends on the recipient’s knowledge),
- **timeliness** (temporal structure),
- **space** (spatial structure).

**Informally Speaking:**
A piece of data carries information if it can impact a recipient’s ability to achieve the objective of some activity within a given context.

**Definition 1.** The amount of information (in a faultless scenario) $\text{info}(E)$ carried by the event $E$ in the context $C$ as measured for a system with the rules of conduct $R$ is

$$\text{info}_{R,C}(E) = \text{cost}[\text{objective}_R(C(E)), \text{objective}_R(C(E) + E)]$$

where the **cost** (weight, distance) is taken according to the ordering of points in the space of objectives.
In our setting, Shannon defined:

**objective:** statistical ignorance of the recipient; statistical uncertainty of the recipient.

**cost:** 
# binary decisions to describe $E$; 
$= -\log P(E)$; $P(E)$ being the probability of $E$.

**Context:** the semantics of data is irrelevant . . .

Self-information for $E_i$: 
info($E_i$) = $-\log P(E_i)$.

Average information: 
$H(P) = -\sum_i P(E_i) \log P(E_i)$

Entropy of $X = \{E_1, \ldots\}$: 
$H(X) = -\sum_i P(E_i) \log P(E_i)$

Mutual Information: 
$I(X; Y) = H(Y) - H(Y|X)$, (faulty channel).

**Theorem 2. (Shannon 1948; Channel Coding)**

In Shannon’s words:

It is possible to **send information** at the **capacity** through the channel with as small a frequency of errors as desired by proper (**long**) encoding. This statement is **not true** for any rate greater than the capacity.

(The maximum codebook size $N(n, \varepsilon)$ for code length $n$ and error probability $\varepsilon$ is asymptotically equal to: 
$N(n, \varepsilon) \sim 2^{nC}$.)

M. Eigen

“The differentiable characteristic of the living systems is Information. Information assures the controlled reproduction of all constituents, thereby ensuring conservation of viability . . . . Information theory, pioneered by Claude Shannon, cannot answer this question . . . in principle, the answer was formulated 130 years ago by Charles Darwin.

Some Fundamental Questions:

- how information is generated and transferred through underlying mechanisms of variation and selection;
- how information in biomolecules (sequences and structures) relates to the organization of the cell;
- whether there are error correcting mechanisms (codes) in biomolecules;
- and how organisms survive and thrive in noisy environments.

Life is a delicate interplay of energy, entropy, and information; essential functions of living beings correspond to the generation, consumption, processing, preservation, and duplication of information.
Beyond Shannon

Participants of the 2005 Information Beyond Shannon workshop realize:

**Time**: When information is transmitted over networks of gene regulation, protein interactions, the associated delay is an important factor. (e.g., timely information exchange in cells may be responsible for bidirectional microtubule-based transport in cells).

**Space**: In molecular interaction networks, spatially distributed components raise fundamental issues of limitations in information exchange since the available resources must be shared, allocated, and re-used.

**Information and Control**: Again in networks involving regulation, signaling, and metabolism, information is exchanged in space and time for decision making, thus timeliness of information delivery along with reliability and complexity constitute the basic objective.

**Semantics**: In many contexts, experimentalists are interested in signals, without precise knowledge of what they represent.

**Dynamic information**: In a complex network in a space-time-control environment (e.g., human brain information is not simply communicated but also processed) how can the consideration of such dynamical sources be incorporated into the Shannon-theoretic model?
Outline Update

1. What is Information?
2. Beyond Shannon Information
3. Information Transfer: Darwin Channel
   - Model: Deletion and Constrained Channels
   - Capacity of the Noisy Constrained Channel
4. Information Discovery in Massive Data: Pattern Matching
5. Information in Structures: Network Motifs
Biomolecular structures, species, and in general biodiversity, have gone through significant metamorphosis over eons through mutation and natural selection, which we model by constrained sequences/channels.

To capture mutation and natural selection we introduce

**Darwin channel**

which is a combination of a noisy deletion/insertion channel and a noisy constrained channel.
Noisy Constrained Channel

1. Binary Symmetric Channel (BSC):
(i) crossover probability $\varepsilon$,
(ii) **constrained set of inputs** (Darwin preselected) that can be modeled by a Markov Process,
(ii) $S_n$ denotes the set of binary constrained sequences of length $n$.

2. Channel Input and Output:
Input: Stationary process $X = \{X_k\}_{k \geq 1}$ supported on $S = \bigcup_{n > 0} S_n$.
Channel Output: **Hidden Markov Process** (HMP)

\[ Z_i = X_i \oplus E_i \]

where $\oplus$ denotes addition modulo 2, and $E = \{E_k\}_{k \geq 1}$, independent of $X$, with $P(E_i = 1) = \varepsilon$ is a Bernoulli process (noise).

**Note**: To focus, we illustrate our results on

\[ S_n = \{(d,k) \text{ sequences}\} \]

i.e., no sequence in $S_n$ contains a run of zeros of length is shorter than $d$ or longer than $k$. Such sequences can model neural spike trains (no two spikes in a short time).
Noisy Constrained Capacity

\(C(\varepsilon)\) – conventional BSC channel capacity \(C(\varepsilon) = 1 - H(\varepsilon)\), where \(H(\varepsilon) = -\varepsilon \log \varepsilon - (1 - \varepsilon) \log (1 - \varepsilon)\).

\(C(S, \varepsilon)\) – noisy constrained capacity defined as

\[
C(S, \varepsilon) = \sup_{X \in S} I(X; Z) = \lim_{n \to \infty} \frac{1}{n} \sup_{X^n_1 \in S_n} I(X^n_1, Z^n_1),
\]

where the suprema are over all stationary processes supported on \(S\) and \(S_n\), respectively. This is an open problem since Shannon.

Mutual information

\[
I(X; Z) = H(Z) - H(Z|X)
\]

where \(H(Z|X) = H(\varepsilon)\).

Thus, we must find the entropy \(H(Z)\) of a hidden Markov process! (e.g., \((d, k)\) sequence can be generated as an output of a \(k\)-th order Markov process).
Hidden Markov Process

1. Let $X = \{X_k\}_{k \geq 1}$ be a $r$th order stationary Markov process over a binary alphabet $\mathcal{A}$ with transition probabilities $P(X_t = a | X_{t-r}^{t-1} = a'_1)$, where $a'_1 \in \mathcal{A}^r$. For $r = 1$, with transition matrix $P = \{p_{ab}\}_{a,b \in \{0,1\}}$.

2. Let $\tilde{X} = 1 \oplus X$. In particular, $Z_i = X_i$ if $E_i = 0$ and $Z_i = \tilde{X}_i$ if $E_i = 1$:

$$P(Z_1^n, E_n) = P(Z_1^n, E_{n-1} = 0, E_n) + P(Z_1^n, E_{n-1} = 1, E_n)$$

$$= P(Z_1^{n-1}, Z_n, E_{n-1} = 0, E_n) + P(Z_1^{n-1}, Z_n, E_{n-1} = 1, E_n)$$

$$= P(E_n) P_X(Z_n \oplus E_n | Z_{n-1}) P(Z_1^{n-1}, E_{n-1} = 0)$$

$$+ P(E_n) P_X(Z_n \oplus E_n | \tilde{Z}_{n-1}) P(Z_1^{n-1}, E_{n-1} = 1)$$
Let
\[ p_n = [P(Z_1^n, E_n = 0), P(Z_1^n, E_n = 1)] \]
and
\[ M(Z_{n-1}, Z_n) = \begin{bmatrix}
(1 - \varepsilon)P_X(Z_n|Z_{n-1}) & \varepsilon P_X(\bar{Z}_n|Z_{n-1}) \\
(1 - \varepsilon)P_X(Z_n|\bar{Z}_{n-1}) & \varepsilon P_X(\bar{Z}_n|\bar{Z}_{n-1})
\end{bmatrix}. \]
Then
\[ p_n = p_{n-1}M(Z_{n-1}, Z_n). \]
and
\[ P(Z_1^n) = p_1M(Z_1, Z_2) \cdots M(Z_{n-1}, Z_n)1^t, \]
where \(1^t = (1, \ldots, 1)\).

Thus, \( P(Z_1^n) \) is a product of random matrices since \( P_X(Z_i|Z_{i-1}) \) are random variables.
Entropy Rate as a Lyapunov Exponent

**Theorem 1 (Furstenberg and Kesten, 1960).** Let $M_1, \ldots, M_n$ form a stationary ergodic sequence and $E[\log^+ ||M_1||] < \infty$ Then

$$\lim_{n \to \infty} \frac{1}{n} E[\log ||M_1 \cdots M_n||] = \lim_{n \to \infty} \frac{1}{n} \log ||M_1 \cdots M_n|| = \mu \quad \text{a.s.}$$

where $\mu$ is called top Lyapunov exponent.

**Corollary 1.** Consider the HMP $Z$ as defined above. The entropy rate

$$h(Z) = \lim_{n \to \infty} E\left[ -\frac{1}{n} \log P(Z^n_1) \right]$$

$$= \lim_{n \to \infty} \frac{1}{n} E\left[ -\log \left( p_1 M(Z_1, Z_2) \cdots M(Z_{n-1}, Z_n) 1^t_t \right) \right]$$

is a top Lyapunov exponent of $M(Z_1, Z_2) \cdots M(Z_{n-1}, Z_n)$.

Unfortunately, it is notoriously difficult to compute top Lyapunov exponents as proved in Tsitsiklis and Blondel. Therefore, in next we derive an explicit asymptotic expansion of the entropy rate $h(Z)$.
We now assume that \( P(E_i = 1) = \varepsilon \to 0 \) is small (e.g., \( \varepsilon = 10^{-12} \) for mutation).

**Theorem 2 (Seroussi, Jacquet and W.S., 2004).** Assume \( r \)th order Markov. If the conditional probabilities in the Markov process \( X \) satisfy

\[
P(a_{r+1}|a_r^r) > 0 \quad \text{IMPORTANT!}
\]

for all \( a_{r+1}^r \in A^{r+1} \), then the entropy rate of \( Z \) for small \( \varepsilon \) is

\[
h(Z) = \lim_{n \to \infty} \frac{1}{n} H_n(Z^n) = h(X) + f_1(P)\varepsilon + O(\varepsilon^2),
\]

where

\[
f_1(P) = \sum_{z_{1}^{2r+1}} P_X(z_{1}^{2r+1}) \log \frac{P_X(z_{1}^{2r+1})}{P_X(\bar{z}_{1}^{2r+1})} = \mathcal{D} \left( P_X(z_{1}^{2r+1}) \parallel P_X(\bar{z}_{1}^{2r+1}) \right),
\]

where \( \bar{z}_{1}^{2r+1} = z_1 \ldots z_r \bar{z}_{r+1} z_{r+2} \ldots z_{2r+1} \). In the above, \( h(X) \) is the entropy rate of the Markov process \( X \), \( \mathcal{D} \) denotes the Kullback-Liebler divergence.
**Examples**

**Example 1.** Consider a Markov process with symmetric transition probabilities $p_{01} = p_{10} = p$, $p_{00} = p_{11} = 1 - p$. This process has stationary probabilities $P_X(0) = P_X(1) = \frac{1}{2}$. Then

$$h(Z) = h(X) + f_1(p)\varepsilon + f_2(p)\varepsilon^2 + O(\varepsilon^3)$$

where

$$f_1(p) = 2(1 - 2p)\log\frac{1 - p}{p}, \quad f_2(p) = -f_1(p) - \frac{1}{2}\left(\frac{2p - 1}{p(1 - p)}\right)^2.$$ 

**Example 2. (Degenerate Case.)** Consider the following Markov process

$$P = \begin{bmatrix} 1 - p & p \\ 1 & 0 \end{bmatrix}$$

where $0 \leq p \leq 1$. Ordentlich and Weissman (2004) proved for this case

$$H(Z) = H(P) - \frac{p(2 - p)}{1 + p}\varepsilon\log\varepsilon + O(\varepsilon)$$

(e.g., $(11\ldots)$ will not be generated by MC, but can be outputed by HMM with probability $O(\varepsilon^\kappa)$).
Recall $I(X; Z) = H(Z) - H(\varepsilon)$. Then by above theorem

$$H(Z) = \mu(P)$$

where $\mu(P)$ is the top Lyapunov exponent of $\{M(\tilde{Z}_i | \tilde{Z}_{i-1})\}_{i>0}$.

It is known (cf. Chen and Siegel, 2004) that the process optimizing the mutual information can be approached by a sequence of Markov probabilities $P^{(r)}$ of the constraint of increasing order.

**Theorem 3.** The noisy constrained capacity $C(S, \varepsilon)$ for a $(d, k)$ constraint through a BSC channel of parameter $\varepsilon$ is given by

$$C(S, \varepsilon) = \lim_{r \to \infty} \sup_{P^{(r)}} \mu(P^{(r)}) - H(\varepsilon)$$

where $P^{(r)}$ denotes the probability law of an $r$-th-order Markov process generating the $(d, k)$ constraint $S$. 
Main Asymptotic Results

We observe (cf. Han and Marcus (2007))

\[ H(Z) = H(P) - f_0(P)\varepsilon \log \varepsilon + f_1(P)\varepsilon + o(\varepsilon) \]

for explicitly computable \(f_0(P)\) and \(f_1(P)\).

Let \(P^{\text{max}}\) be the maxentropic maximizing \(H(P)\). Then

\[ C(S, \varepsilon) = C(S) - (1 - f_0(P^{\text{max}}))\varepsilon \log \varepsilon + (f_1(P^{\text{max}}) - 1)\varepsilon + o(\varepsilon) \]

where \(C(S)\) is known capacity of a noiseless channel.

**Example:** For \((d, k)\) sequences, we can prove:

(i) for \(k \leq 2d\)

\[ C(S, \varepsilon) = C(S) + A \cdot \varepsilon + O(\varepsilon^2 \log \varepsilon) \]

(ii) For \(k > 2d\)

\[ C(S, \varepsilon) = C(S) + B \cdot \varepsilon \log \varepsilon + O(\varepsilon), \]

where \(A\) & \(B\) are computable constants (cf. also Han and Marcus (2007)).
1. What is Information?
2. Information Transfer: Darwin Channel
3. Information Discovery in Massive Data: Pattern Matching
   - Classification of Pattern Matching Problems
   - Statistical Significance
   - Finding Weak Signals in Biological Data
4. Information in Structures: Network Motifs
Pattern Matching

Let $W = w_1 \ldots w_m$ and $T$ be strings over a finite alphabet $A$.

Basic question: **how many times $W$ occurs in $T$.**

Define $O_n(W)$ — the number of times $W$ occurs in $T$, that is,

$$O_n(W) = \# \{ i : T_{i-m+1}^i = W, m \leq i \leq n \}.$$ 

**Generalized String Matching:** A set of patterns is given, that is,

$$W = (W_0, W_1, \ldots, W_d), \quad W_i \in A^{m_i}$$

where $W_i$ itself for $i \geq 1$ is a subset of $A^{m_i}$ (i.e., a set of words of a given length $m_i$). The set $W_0$ is called the forbidden set.

**Three cases to be considered:**

- $W_0 = \emptyset$ — the number of patterns from $W$ occurring in the text.
- $W_0 \neq \emptyset$ — the number of $W_i, i \geq 1$ pattern occurrences under the condition that no pattern from $W_0$ occurs in the text.
- $W_i = \emptyset, i \geq 1, W_0 \neq \emptyset$ — e.g., $(d, k)$ sequences.
Probabilistic Sources

**Memoryless Source**
The text is a realization of an independently, identically distributed sequence of random variables (i.i.d.), such that a symbol \( s \in A \) occurs with probability \( P(s) \).

**Markovian Source**
The text is a realization of a stationary Markov sequence of order \( r \), that is, probability of the next symbol occurrence depends on \( r \) previous symbols.

**Basic Thrust of our Approach**
When searching for over-represented or under-represented patterns we must assure that such a pattern is not generated by randomness itself (to avoid too many false positives).
Some statistical tools are used to characterize underrepresented and overrepresented patterns. We illustrate it on $O_n(\mathcal{W})$.

**Z-scores**

$$Z(\mathcal{W}) = \frac{\mathbb{E}[O_n] - O_n(\mathcal{W})}{\sqrt{\text{Var}(O_n(\mathcal{W}))}}$$

i.e., how many standard deviations the observed value $O_n(\mathcal{W})$ is away from the mean.

This score makes sense only if one can prove that $Z$ satisfies (at least asymptotically) the Central Limit Theorem (CLT), that is, $Z$ is normally distributed.

**p-values**

$$pval(r) = P(O_n(\mathcal{W}) > \mathbb{E}[O_n] + x \sqrt{\text{Var}[O_n]})$$

$p$ values are used for very rare occurrences, far away from the mean.

In order to compute $p$ values one must apply either Moderate Large deviation (MLD) or Large Deviations (LD) results.
Let \( P(O_n \geq n\alpha + x\sigma\sqrt{n}) \)

**Central Limit Theorem (CLT)** – valid only for \( x = O(1) \).

**Moderate Large Deviations (MLD)** – valid for \( x \to \infty \) but \( x = o(\sqrt{n}) \).

**Large Deviations (MLD)** – valid for \( x = O(\sqrt{n}) \).
**Z-scores and p values for A.thaliana**

Table 1: Z score vs p-value of tandem repeats in *A.thaliana*.

<table>
<thead>
<tr>
<th>Oligomer</th>
<th>Obs.</th>
<th>p-val (large dev.)</th>
<th>Z-sc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AATTGGCGGG</td>
<td>2</td>
<td>$8.059 \times 10^{-4}$</td>
<td>48.71</td>
</tr>
<tr>
<td>TTTGTACCA</td>
<td>3</td>
<td>$4.350 \times 10^{-5}$</td>
<td>22.96</td>
</tr>
<tr>
<td>ACGGTTCAC</td>
<td>3</td>
<td>$2.265 \times 10^{-6}$</td>
<td>55.49</td>
</tr>
<tr>
<td><strong>AAGACGGTT</strong></td>
<td>3</td>
<td>$2.186 \times 10^{-6}$</td>
<td>48.95</td>
</tr>
<tr>
<td>ACGACGCTT</td>
<td>4</td>
<td>$1.604 \times 10^{-9}$</td>
<td>74.01</td>
</tr>
<tr>
<td>ACGCTTTGG</td>
<td>4</td>
<td>$5.374 \times 10^{-10}$</td>
<td>84.93</td>
</tr>
<tr>
<td>GAGAAGACG</td>
<td>5</td>
<td>$0.687 \times 10^{-14}$</td>
<td>151.10</td>
</tr>
</tbody>
</table>

Remark: p values were computed using large deviations results of Regnier and W.S. (1998), and Denise and Regnier (2001) as we discuss below.
Some Theory: Language Based Approach


**Language Approach:** Language $\mathcal{L}$ a collection of words, and its generating function is

$$ L(z) = \sum_{u \in \mathcal{L}} P(u) z^{|u|} $$

where $P(u)$ is the probability $u$ occurrence, $|u|$ is the length of $u$.

**Autocorrelation Set and Polynomial:**
Given a pattern $\mathcal{W}$, we define the autocorrelation set $S$ as:

$$ S = \{ w_{k+1}^m : w_1^k = w_{m-k+1}^m \}, \quad w_1^k = w_{m-k+1}^m $$

and $\mathcal{W}W$ is the set of positions $k$ satisfying $w_1^k = w_{m-k+1}^m$.

The generating function of $S$ is denoted as $S(z)$ and we call it the autocorrelation polynomial.

$$ S(z) = \sum_{k \in \mathcal{W}W} P(w_{k+1}^m) z^{m-k}. $$
Language $\mathcal{T}_r$

$\mathcal{T}_r$ – set of words that contains exactly $r \geq 1$ occurrences of $\mathcal{W}$. Define

$$T_r(z) = \sum_{n \geq 0} \Pr\{O_n(\mathcal{W}) = r\} z^n, \quad T(z, u) = \sum_{r=1}^{\infty} T_r(z) u^r.$$ 

(i) We define $\mathcal{R}$ as the set of words containing only one occurrence of $\mathcal{W}$, located at the right end; e.g., for $\mathcal{W} = aba$, $ccaba \in \mathcal{R}$.

(ii) We also define $\mathcal{U}$ as

$$\mathcal{U} = \{u : \mathcal{W} \cdot u \in \mathcal{T}_1\}$$

that is, a word $u \in \mathcal{U}$ if $\mathcal{W} \cdot u$ has exactly one occurrence of $\mathcal{W}$ at the left end of $\mathcal{W} \cdot u$; e.g., $bba \in \mathcal{U}$ but $ba \notin \mathcal{U}$.

(iii) Let $\mathcal{M}$ be the language:

$$\mathcal{M} = \{u : \mathcal{W} \cdot u \in \mathcal{T}_2 \text{ and } \mathcal{W} \text{ occurs at the right of } \mathcal{W} \cdot u\},$$

that is, $\mathcal{M}$ is a language such that $\mathcal{W} \mathcal{M}$ has exactly two occurrences of $\mathcal{W}$ at the left and right end of a word from $\mathcal{M}$; e.g., $ba \in \mathcal{M}$ since $ababa \in \mathcal{T}_2$. 
Lemma 1. The language $T$ satisfies the fundamental equation:

$$T = R \cdot M^* \cdot U.$$ 

Notably, the language $T_r$ can be represented for any $r \geq 1$ as follows:

$$T_r = R \cdot M^{r-1} \cdot U,$$

and

$$T_0 \cdot W = R \cdot S.$$ 

Here, by definition $M^0 := \{\epsilon\}$ and $M^* := \bigcup_{r=0}^{\infty} M^r$.

Example: Let $W = TAT$. The following string belongs $T_3$:

$$\underbrace{R}_{M} \underbrace{CCTAT}_{M} \underbrace{AT}_{M} \underbrace{GATAT}_{M} \underbrace{GGA}_{U}.$$
Main Results – Exact

Theorem 4. (i) The languages $\mathcal{M}, \mathcal{U}$ and $\mathcal{R}$ satisfy:

$$\bigcup_{k \geq 1} \mathcal{M}^k = \mathcal{A}^* \cdot \mathcal{W} + S - \{\epsilon\},$$

$$\mathcal{U} \cdot \mathcal{A} = \mathcal{M} + \mathcal{U} - \{\epsilon\},$$

$$\mathcal{W} \cdot \mathcal{M} = \mathcal{A} \cdot \mathcal{R} - (\mathcal{R} - \mathcal{W}),$$

where $\mathcal{A}^*$ is the set of all words, $+$ and $-$ are disjoint union and subtraction of languages.

(ii) The generating functions $T_r(z)$ and $T(z, u)$ are

$$T_r(z) = R(z) M^{r-1} U(z), \quad r \geq 1$$

$$T(z, u) = R(z) \frac{u}{1 - u M(z)} U(z)$$

$$T_0(z) P(\mathcal{W}) = R(z) S(z)$$

where

$$M(z) = 1 + \frac{z - 1}{D(z)}, \quad U(z) = \frac{1}{D(z)}, \quad R(z) = z^m P(\mathcal{W}) \frac{1}{D(z)}.$$

with $D(z) = (1 - z)S(z) + z^m P(\mathcal{W})$. 
Main Results: Asymptotics

**Theorem 5.** (i) **Moments.** The expectation satisfies, for \( n \geq m \):

\[
E[O_n(W)] = P(W)(n - m + 1),
\]

while the variance is

\[
\text{Var}[O_n(W)] = nc_1 + c_2.
\]

where \( c_1, c_2 \) are explicit constants.

(ii) **CLT:** Case \( r = EO_n + x\sqrt{\text{Var}O_n} \) for \( x = O(1) \). Then:

\[
\Pr\{O_n(W) = r\} = \frac{1}{\sqrt{2\pi c_1 n}} e^{-\frac{1}{2}x^2} \left( 1 + O\left(\frac{1}{\sqrt{n}}\right)\right).
\]

(iii) **Large Deviations:** Case \( r = (1 + \delta)EO_n \). Let \( a = (1 + \delta)P(W) \), then

\[
\Pr\{O_n(W) \sim (1 + \delta)EO_n\} = \frac{e^{-(n-m+1)I(a) + \delta_a}}{\sigma_a \sqrt{2\pi(n - m + 1)}}
\]

where \( I(a) = a\omega_a + \rho(\omega_a) \) and \( \delta_a, \omega_a, \rho(\omega_a) \) are constants, and \( \rho \) is the smallest root of \( D(z) = 0 \).
Denise and Regnier (2002) observed that in biological sequence whenever a word is overrepresented, then its subwords and proximate words are also likely to be overrepresented (the so called artifacts). Example: if $W_1 = AATAAA$, then $W_2 = ATAAA$ is also overrepresented.

**New Approach:**

Once a dominating signal has been detected, we look for a weaker signal by comparing the number of observed occurrences of patterns to the conditional expectations not the regular expectations.

In particular, using the methodology presented above Denise and Regnier (2002) were able to prove that

$$E[O_n(W_2)|O_n(W_1) = k] \sim \alpha n$$

provided $W_1$ is overrepresented, where $\alpha$ can be explicitly computed (often $\alpha = P(W_2)$ is $W_1$ and $W_2$ do not overlap).
Beaudouin et al. (2000) studied several variants of the well known AAUAAA polyadenylation signal in mRNA of human genes. To avoid artifacts Beaudouin et al cancelled all sequences where the overrepresented hexamer was found.

Using our approach Denise and Regnier (2002) discovered/eliminated all artifacts and found new signals in a much simpler and reliable way.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AAUAAA</td>
<td>3456</td>
<td>1</td>
<td>363.16</td>
<td>167.03</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAAUAA</td>
<td>1721</td>
<td>2</td>
<td>363.16</td>
<td>71.25</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUAAAA</td>
<td>1530</td>
<td>3</td>
<td>363.16</td>
<td>61.23</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UUUUUUU</td>
<td>1105</td>
<td>4</td>
<td>416.36</td>
<td>33.75</td>
<td>8</td>
<td>373.30</td>
<td>37.87</td>
<td>2</td>
</tr>
<tr>
<td>AUAAAU</td>
<td>1043</td>
<td>5</td>
<td>373.23</td>
<td>34.67</td>
<td>6</td>
<td>1529.15</td>
<td>12.43</td>
<td>4078</td>
</tr>
<tr>
<td>AAAAUA</td>
<td>1019</td>
<td>6</td>
<td>363.16</td>
<td>34.41</td>
<td>7</td>
<td>848.76</td>
<td>5.84</td>
<td>420</td>
</tr>
<tr>
<td>UAAAAU</td>
<td>1017</td>
<td>7</td>
<td>373.23</td>
<td>33.32</td>
<td>9</td>
<td>780.18</td>
<td>8.48</td>
<td>211</td>
</tr>
<tr>
<td>AUUAAA</td>
<td>1013</td>
<td>1</td>
<td>373.23</td>
<td>33.12</td>
<td>10</td>
<td>385.85</td>
<td>31.93</td>
<td>3</td>
</tr>
<tr>
<td>AUAAGA</td>
<td>972</td>
<td>9</td>
<td>184.27</td>
<td>58.03</td>
<td>4</td>
<td>593.90</td>
<td>15.51</td>
<td>34</td>
</tr>
<tr>
<td>UAAAAU</td>
<td>922</td>
<td>10</td>
<td>373.23</td>
<td>28.41</td>
<td>13</td>
<td>1233.24</td>
<td>–8.86</td>
<td>4034</td>
</tr>
<tr>
<td>UAAAAA</td>
<td>922</td>
<td>11</td>
<td>363.16</td>
<td>29.32</td>
<td>12</td>
<td>922.67</td>
<td>9.79</td>
<td>155</td>
</tr>
<tr>
<td>UAAAAA</td>
<td>863</td>
<td>12</td>
<td>373.23</td>
<td>25.35</td>
<td>15</td>
<td>374.81</td>
<td>25.21</td>
<td>4</td>
</tr>
<tr>
<td>CAAAUA</td>
<td>847</td>
<td>13</td>
<td>185.59</td>
<td>48.55</td>
<td>5</td>
<td>613.24</td>
<td>9.44</td>
<td>167</td>
</tr>
<tr>
<td>AAAAAA</td>
<td>841</td>
<td>14</td>
<td>353.37</td>
<td>25.94</td>
<td>14</td>
<td>496.38</td>
<td>15.47</td>
<td>36</td>
</tr>
<tr>
<td>UAAAAUA</td>
<td>805</td>
<td>15</td>
<td>373.23</td>
<td>22.35</td>
<td>21</td>
<td>1143.73</td>
<td>–10.02</td>
<td>4068</td>
</tr>
</tbody>
</table>
Outline Update

1. What is Information?
2. Information Transfer: Darwin Channel
3. Information Discovery in Massive Data: Pattern Matching
4. Information in Structures: Network Motifs
   - Biological Networks
   - Finding Biologically Significant Structures
Protein Interaction Networks

- **Molecular Interaction Networks**: Graph theoretical abstraction for the organization of the cell

- **Protein-protein interactions** (PPI Network)
  - Proteins signal to each other, form complexes to perform a particular function, transport each other in the cell...
  - It is possible to detect interacting proteins through high-throughput screening, small scale experiments, and *in silico* predictions

![Undirected Graph Model](image)

*S. Cerevisiae* PPI network

(Jeong et al., *Nature*, 2001)
Modularity in PPI Networks

- A **functionally modular group** of proteins (e.g. a protein complex) is likely to induce a **dense subgraph**

- **Algorithmic approaches** target identification of dense subgraphs

- An important problem: **How do we define dense?**
  - Statistical approach: What is **significantly** dense?

RNA Polymerase II Complex  |  Corresponding induced subgraph
Significance of Dense Subgraphs

- A subnet of \( r \) proteins is said to be \( \rho \)-dense if the number of interactions, \( F(r) \), between these \( r \) proteins is \( \geq \rho r^2 \), that is,

\[
F(r) \geq \rho r^2
\]

- What is the expected size, \( R_\rho \), of the largest \( \rho \)-dense subgraph in a random graph?
  - Any \( \rho \)-dense subgraph with larger size is statistically significant!
  - Maximum clique is a special case of this problem (\( \rho = 1 \))

- \( G(n, p) \) model
  - \( n \) proteins, each interaction occurs with probability \( p \)
  - Simple enough to facilitate rigorous analysis

- Piecewise \( G(n, p) \) model
  - Captures the basic characteristics of PPI networks

- Power-law model
Largest Dense Subgraph on $G(n, \rho)$

**Theorem 6.** If $G$ is a random graph with $n$ nodes, where every edge exists with probability $\rho$, then

$$\lim_{n \to \infty} \frac{R_\rho}{\log n} = \frac{1}{H_p(\rho)} \quad (pr.),$$

where

$$H_p(\rho) = \rho \log \frac{\rho}{p} + (1 - \rho) \log \frac{1 - \rho}{1 - p}$$

denotes divergence. More precisely,

$$P(R_\rho \geq r_0) \leq O \left( \frac{\log n}{n^{1/H_p(\rho)}} \right),$$

where

$$r_0 = \frac{\log n - \log \log n + \log H_p(\rho)}{H_p(\rho)}$$

for large $n$. 
Proof

- \( X_{r,\rho} \): number of subgraphs of size \( r \) with density at least \( \rho \)
  \[ X_{r,\rho} = |\{U \subseteq V(G) : |U| = r \land |F(U)| \geq \rho r^2\}| \]

- \( Y_r \): number of edges induced by a set \( U \) of \( r \) vertices
  \[
  E[X_r] = \binom{n}{r} P(Y_r \geq \rho r^2) \\
  P(Y_r \geq \rho r^2) \leq \binom{r^2}{\rho r^2}(r^2 - \rho r^2)p^{\rho r^2}(1 - p)^{r^2 - \rho r^2} 
  \]

- **Upper bound:** By the first moment method:
  \[
  P(R_\rho \geq r) \leq P(X_{r,\rho} \geq 1) \leq E[X_{r,\rho}] 
  \]

  - Plug in Stirling’s formula for appropriate regimes

- **Lower bound:** To use the second moment method, we have to account for dependencies in terms of nodes and existing edges

  - Use Stirling’s formula, plug in continuous variables for range of dependencies
Piecewise $G(n, p)$ Model

- Few proteins with many interacting partners, many proteins with few interacting partners

Captures the basic characteristics of PPI networks, where $p_l < p_b < p_h$.

- $G(V, E), V = V_h \cup V_l$ such that $n_h = |V_h| \ll |V_l| = n_l$

$$P(uv \in E(G)) = \begin{cases} 
p_h & \text{if } u, v \in V_h \\
p_l & \text{if } u, v \in V_l \\
p_b & \text{if } u \in V_h, v \in V_l \text{ or } u \in V_l, v \in V_h. \end{cases}$$

- If $n_h = O(1)$, then $P(R_\rho \geq r_1) \leq O\left(\frac{\log n}{n^{1/H_{p_l}(\rho)}}\right)$ where

$$r_1 = \frac{\log n - \log \log n + 2n_h \log B + \log H_{p_l}(\rho) - \log e + 1}{H_{p_l}(\rho)}$$

and $B = (p_b(1 - p_l))/p_l + (1 - p_b)$. 
SiDES

• An algorithm for identification of Significantly Dense Subgraphs (SiDES)
  – Based on Highly Connected Subgraphs algorithm (Hartuv & Shamir, 2000)
  – Recursive min-cut partitioning heuristic
  – We use statistical significance as stopping criterion
Behavior of Largest Dense Subgraph Across Species

Number of nodes vs. Size of largest dense subgraph for PPI networks belonging to 9 Eukaryotic species
Behavior of Largest Dense Subgraph w.r.t Density

S. cerevisiae

H. sapiens

Density threshold vs. Size of largest dense subgraph for Yeast and Human PPI networks
Performance of SiDES

- **Biological relevance** of identified clusters is assessed with respect to Gene Ontology (GO)
  - Find GO Terms that are significantly enriched in each cluster

- **Quality** of the clusters
  - For GO Term \( t \) that is significantly enriched in cluster \( C \), let \( T \) be the set of proteins associated with \( t \)

\[
\text{specificity} = 100 \times \frac{|C \cap T|}{|C|} \\
\text{sensitivity} = 100 \times \frac{|C \cap T|}{|T|}
\]

<table>
<thead>
<tr>
<th></th>
<th>SiDES</th>
<th>MCODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specificity (%)</td>
<td>43.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>2.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Comparison of SiDES with MCODE (Bader & Hogue, 2003) on yeast PPI network derived from DIP and BIND databases
Performance of SiDES

Correlation

SiDES: 0.22
MCODE: -0.02

SiDES: 0.27
MCODE: 0.36