数理解析研究所講究録
第 1698 巻 2010 年 104-108

2, How to correct errors in biological information processing -A suggestion from state of the art error correcting codes-

樺島祥介

1 背景

生物はノイズの強い環境においても情報を正確に伝達しているようにみえる(ノイズ耐性があるようにみえる)。よって、何かしらのエラー訂正の仕組みがあると考えることができる。本講演では、確率的なモデル化によりエラー訂正の方法を議論する数理的な枠組みを紹介し、続いて生物の特性と照らし合わせて生物学的にあり得るエラー訂正の仕組みについて考察する。今回登場するのは誤り訂正符号（Error correcting codes : ECCs）のシステムである。符号とは情報に何らかの機能を持たせるために変換操作を施すことであり、復号とは変換された表現を元の表現に戻すことを意味する。情報の劣化に対する耐性の付与を目的とした符号のことを特に誤り訂正符号と呼ぶ。

2 誤り訂正符号 Error correcting codes : ECCs

まずは確率的なモデル化によりエラー訂正を数学的に扱うシャノンの理論を紹介する。話を簡単化するために情報は 2 つのアルファベット 0, 1 （bit 値）を並べたベクトルで表現されるとし、劣化過程としては与えられた情報（ベクトル）の各要素が独立に確率 p で 0 が 1 に、また、1 が 0 に反転する binary symmetric channel (BSC)を考える。ノイズがある場合、誤り訂正の仕組みがないと情報を正しく伝送することはできない。誤り訂正符号のポイントは、K bit の情報を長くして符号化することにある。直感的には、情報表現にわざと無駄を加えることにより仮に情報が劣化してもこの無駄を手がかりとして元に戻すことができる機能を作り出すのである。このメカニズムを数学的に記述するため、無駄具合を code rate（R）で表す（Fig.1, ＊R=K/N 情報の濃度　K:情報、N:情報＋無駄)。つまり、R が大きい（無駄が少ない）と通信のコストが下がるが、エラーの割合が増えることになる。一方 R が小さい（無駄が多い）とエラーの割合が減るが、通信のコストが上がってしまう。つまり、通信のコストとエラー率の関係はトレードオフの関係になる。シャノンはこのトレードオフの問題を数学的に考察し、K と N が無限大の極限を考えた場合、この無駄具合が

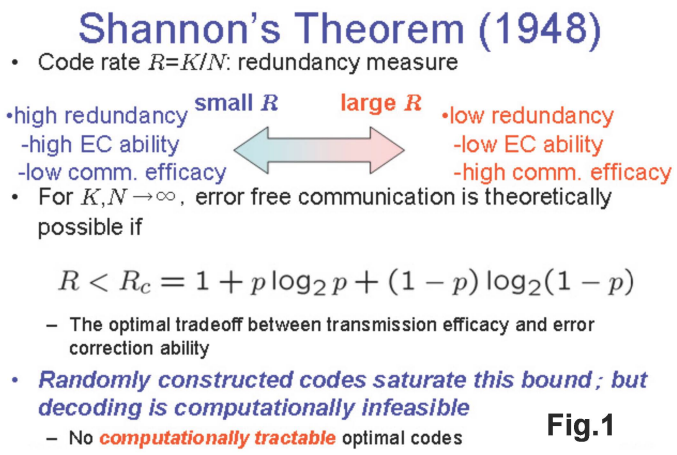


Fig.1

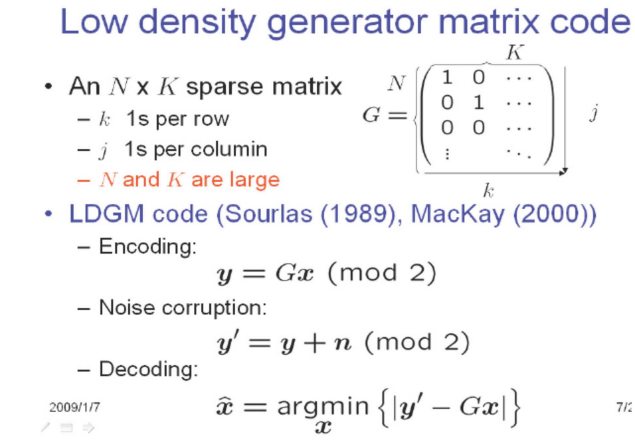


Fig.2

あるノイズによって決まるクリティカルな rate（Rc）よりも小さいときには、理論的にエラーフリーな通信が可能であることを示した(Fig.1)。これは、スケールを大きくすると（N, K を∞に持って行くと）質的に異なる機能が現れてくるという点で、スケール変換の話に似ているかもしれない。ただし、シャノンの考察した符号は原理的には可能だが符号化、復号に bit 長に関して指数関数的に増大する計算コストが必要なためである実際的ではない（現実：リアルタイム通信）。

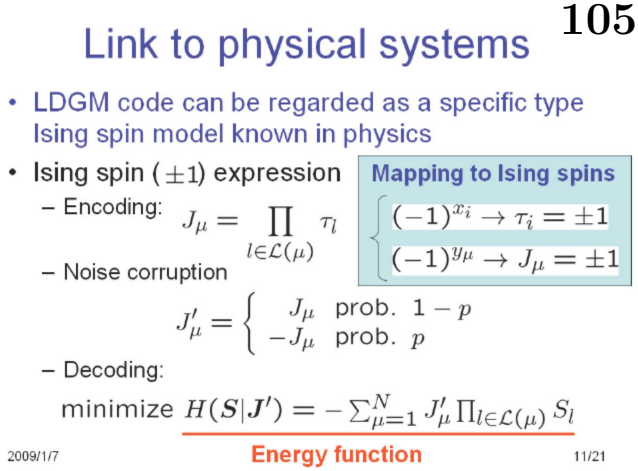


Fig.3

3．Low density generator matrix （LDGM）符号

そのため、シャノンが潜在的な能力の高さを示した“長い符号”は符号研究の中では早期に見切りをつけられ、実際的な符号の研究は長らくの間、代数学の知見に基づいた“短い符号”を中心に発展してきた。ところが、90 年代半ば、疎なランダム行列（疎行列）に基づいた“長い符号”が近似的な復号アルゴリズムを利用することで実際的な計算コストで極めて優れた誤り訂正能力を示すことが実証された。このことを契機に研究の流れが、再度、劇的に変化し現時点では誤り訂正符号の主な研究潮流は“長い疎行列符号”に完全に取って代わられている。ここでは、工学的には必ずしも優れた符号とはいえないが生物学的な実現のしやすさという観点から、最も素朴な疎行列符号と位置づけられる low density generator matrix code を取り上げる。

情報は 0, 1 で表現されると考え、N, K は十分に大きいという仮定の下、線形変換により K bit の情報を N bit の符号語に符号化する方法を考える（Fig.2）。そのために、1 行あたりに 1 がある個数が k 個、1 列あたりに 1 がある個数が j 個の拘束条件をつけた N×K 疎行列(sparse matrix)である generator matrix(G)をランダムにつくる。符号化は、送りたい情報 x に、この G を左から掛けることで行われる(Fig2;Encoding 参照)。ただし、以下、演算は 2 を法として行う。BSC による情報の劣化過程は符号語に反転が生じた成分は 1 を反転が生じなかった成分は 0 としたノイズベクトルを足す操作であると考えることができる（noise corruption 参照）。復号は、受け取った情報 y'に対して Gx から、｜y'- Gx｜を最小化するような x を探し出すことで行われる（逆問題を解く）。このように情報を符号化・復号する仕組みを Low density generator matrix (LDGM) 符号と呼ぶ。LDGM 符号は K よりも十分小さい状況で k を十分大きくするとシャノンが示した誤り訂正に関する最良のトレードオフを 達成することが明らかにされている。

次に、LDGM を用いた誤り訂正符号を統計力学の枠組みで理解することを考える。具体的には LDGM 符号の復号をある特殊な Ising spin model の問題にマップする。そのために、今までの 0, 1 の情報表現を-1, +1 に置き換える。この場合、-1, +1 のかけ算は 0, 1 の足し算に相当するなど、Encoding, Noise corruption は Fig.3 のように置き換わる。ここで難しいのは decoding であるが、これも送られてきた情

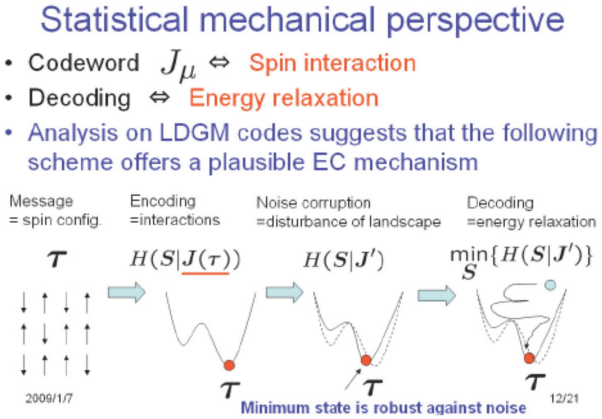


Fig.4

報と受け取った情報の差を最小化することで求まる。このことは、Fig.3 の decoding の式で表わされるエネルギー状態を最小化することと同義になり、統計物理学者の得意分野となる。この場合 codeword は spin interaction の強さを表すものになり、decoding はその spin interaction で決まるようなエネルギー関数の energy relaxation と扱うことができる（Fig.4）。つまり、送りたい情報は微小磁石（Ising spin）の並びと考え（τ）、encoding はこの磁石の並びを基底状態としたエネルギー関数を作ることだと考えることができる（$H(S|J(\tau))$）。しかしこの時、noise corruption（$H(S|J')$）により、エネルギー関数の形(エネルギーランドスケイプ）が変わってしまう。ところが、１つの spin interaction を構成するスピン数（k）がある程度大きい場合に統計力学的な解析を行うと、ノイズの影響による基底状態のずれは小さいことが示され、基底状態はノイズに対してロバストであることが結論付けられる。つまり、このように情報を表現し復号すればノイズに対してロバストな情報伝達が可能になる。

このようにして LDGM 符号を解釈すると、生物の情報伝達システムも同様の誤り訂正符号の仕組みを利用している可能性が見えてくる。通常のコンピュータを用いると LDGM の符号化は容易である一方で、残念ながら復号に要する計算量が膨大になる。ところが、工学的には解決の難しいこの計算困難の問題が生物では生じない。なぜなら、ミクロな世界の低エネルギー状態は＜安定化するまで（解が求まるまで）システムを自然に放っておけば勝手に＞求まってしまうからである。生物システムはこうした“自然法則が行う計算の特性”を上手く利用している可能性がある。

ホップフィールドはこの可能性の一つとして連想記憶模型という記憶（脳）のモデルを考えた（Fig. 5）。この場合、神経には相互作用があり、経験によって強化されている（ヘッブ則）。この状況下で記憶した情報の想起に energy function の考え方をあてはめることが出来る。もう一つの可能性としてはセントラルドグマが挙げられる。この場合のメッセージはタンパク質の３次元構造である。しかしながら、生物はその３次元構造をそのまま扱うのではなく DNA, RNA にコー

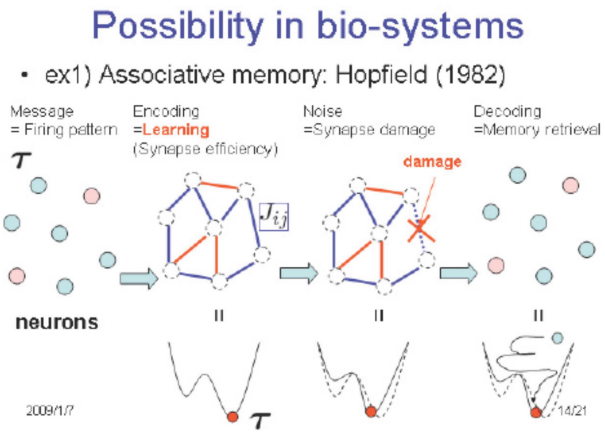


Fig.5

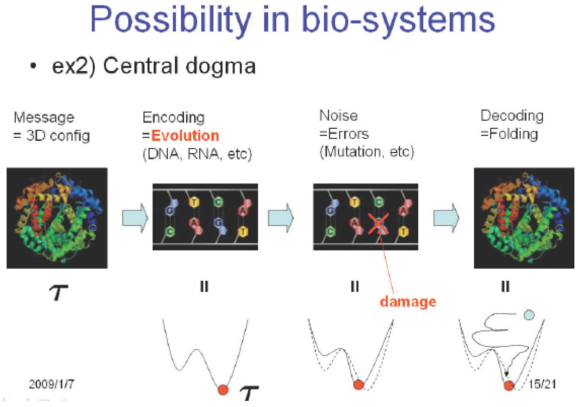


Fig.6

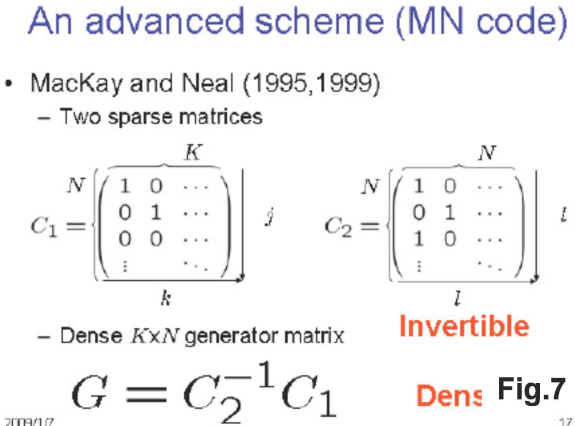
ドし必要に応じてその都度タンパク質を作り出すというまどろっこしい情報伝達方法を採用している。その理由の一つは、そうしておけばノイズ（遺伝子変異）が生じてもエネルギーランドスケイプは大して変わらないので正確な情報伝達が可能になるから、ではないか（Fig.6）ということが推察される。この場合相互作用の強さは進化で獲得したものと考えることができる。

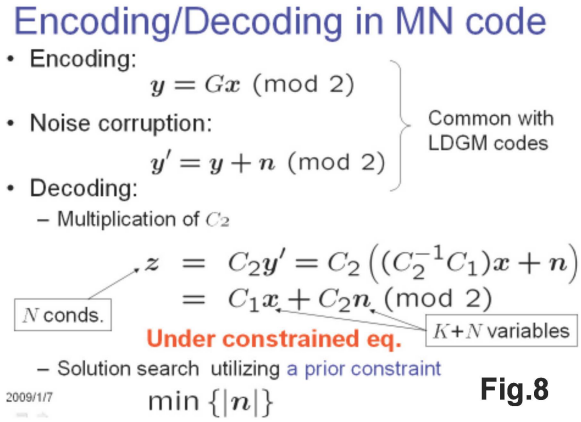
４．ここまでのまとめ

最近の error correcting codes を統計力学的に眺めてみると encoding はメッセージを相互作用に埋め込む作業、decoding は基底状態の探索の作業と見なすことが出来る。そして、このシステムは biological なシステムにおいても鍵になっているのではないだろうか。

５．暗号化

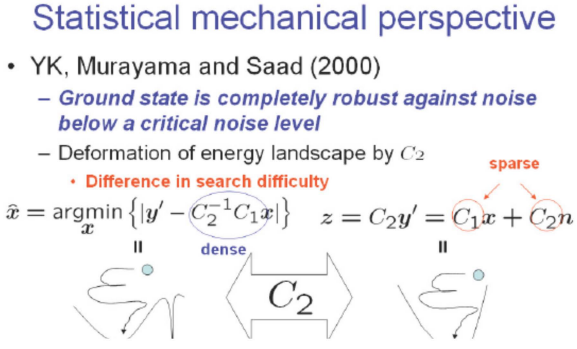
続いて LDGM 符号をさらに発展させ、情報を暗号化する方法について紹介する。先ほどの LDGM 符号では符号化の際、演算子として C1 行列をかけたが、暗号と絡めるためもう一つ疎な正則行列（C2）を用意する。具体的には $C2^{-1}C1$ を LDGM 符号における generator matrix として用いる（Fig.7）。C2 はもともと疎行列であるので $C2^{-1}$ は密行列となる。この点を除いて encoding, noise corruption に関してはこれまでと同様である(Fig.8)。ここでのポイントは decoding にある。$C2^{-1}C1$ は密なマトリックスなので、そのままの形で





復号するのは非常に難しい。しかし、C2 を知っていると受信した信号 y'に C2 を掛けることで疎行列の表現に変換することができる。復号で必要なのは z（N 個の情報）から x, n（K+N（ノイズ）個の情報）を求めることである。これ自体は不定方程式なのでそのままでは解くことはできない。しかし、現実にはノイズはそれほど強くないはずであり（前提条件）、ノイズベクトル n で 1 が立っている成分の個数は少ないはずである。そこで、復号では Fig. 8 の疎行列に基づく不定方程式を制約条件として $\min\{|n|\}$ という条件を解けばよい。この問題はある種の近似アルゴリズムを用いることで効率的に解くことができるのである。

数学的な条件のみに目を向けると、C1 をそのまま送り、解くということと、$C2^{-1}$ をかけて情報を送り C2 をかけて元に戻して解くということは等価である。しかし、物理的に考えるとこの時エナ



ジーランドスケイプは変わってしまうため、最終的に基底状態に到達するための引き込み領域が変化することになる（Fig.9）。工学的には代数的に暗号が破られないように（簡単に符号を解かれない（因数分解されない）ように）、encoding の際に密行列 D をさらにかける（Fig.10）。C2, D を知っていれば簡単に復号することができ、送られてきた正確な情報を得ることができる。ちなみに、この暗号化の仕方は、最終的に安定な状態は変更しないが、安定な状態に至るまでのエネルギーを減少させるという点で、分子シャペロン(タンパク質のフォールディングを補助するタンパク質)の役割を彷彿とさせる。

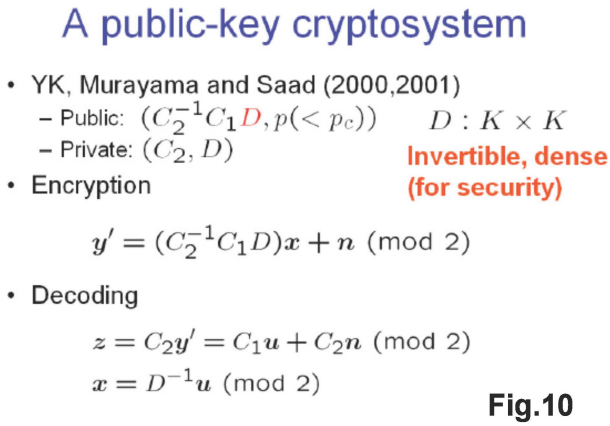


Fig.10

質問

Q1. C2 から $C2^{-1}$ を解くのは難しいのではないか？

A1. 1 回作って、後は使い回し。

Q2. 偶発的に（他者が）解いてしまう可能性は？

A2. それはある。

Q3. スピングラスの理論を使っているのか？

A3. そうだ。

Q4. 基底状態を探索するとき local minimum を探してしまう可能性は？

A4. N を無限大にとばした後は P が十分小さければ大丈夫、それは P に依存

Q5. decoding の問題を最小値化問題に変えたのは解った。ではそれを速く解ける理由は？

A5. （error が少なければ）エナジーランドスケイプ（を解くことが）が簡単だから。

Q6. 最小値を探索する方法のスキームはあるか？

A6. 一般原理はあるが、この性質に合うように組む。（自然界なら放っておけばよい）

Q7. 生物へのアナロジーはタンパク質より分化と繋げたらどうだろう？

→分化の過程で新たなタンパク質が出来ることで相互作用が強化されると考えられるのでは？

A7. 発生の場合エネルギーが定義できないのでは？アトラクターと考えれば悪くない（概念）？

Q8. 計算が何時終わったとするのか？

A8. 前の値と次の値を比較して変わってなければ終わったとする。

→ちょっと難しい話。細かいアドバイスあり。

Q9. 何故 D を入れたのか？

A9. $C2^{-1}$ だけでは解かれる可能性がある。そこで、行列の因数分解を難しくするために入れた。