情報理論と生物学

京都大学理学部２年　市野 悠

2005 年 7 月 9 日

1　はじめに

今回の特集を「バイオインフォマティックス特集」としなかったのには理由があります．「元会計が言ったから」ではなくて（笑），なるべく広く「生物特集」にしたかったからなのです．そんな中，まったくバイオインフォマティックスと関係ない（こともない）記事を書いてみようと思い立ちました．生物門外漢の学部生の話ですので，生物学の常識を逸脱しているかもしれません．是非コメントをいただきたいと思います．

バイオインフォマティックスは，主に計算機を使って生物をミクロな視点から解析しようという試みと言えるでしょうが，その本質は何でしょうか？飯田さんの記事にもありますが，ゲノムをすべて文字列として眺めていても何も分かりません．その暗号解読が必要になります．一方，生命によって作られた暗号には，無駄があります．例えばヒトとチンパンジーの違いを調べたいのに，遺伝子レベルでは 1.7% の違いしかないので，それ以外の部分は研究に直接必要ないでしょう．

これを情報理論と対応させてみましょう．かつて Bell 研究所全盛期（じゃなくて通信業界最盛期）には，いかに回線のノイズ（雑音）を抑えるかということが最重要課題でした．そしてそれに付随する基礎研究，つまり「ノイズとはなんぞや？」という問いに対して優秀な数学者や物理学者が動員されていました．確か研究のおまけでノーベル賞も出たと記憶していますが，とにかく情報理論の基礎がこの時代に築かれたわけです．

電話をかけている自分を想像してみましょう．最近の固定電話どうしでは聞き取れるノイズはほとんどありませんね．しかし忌まわしき携帯電話とかなんとかいう文明の汚染物質（違）から電話がかかってきたとき，ひどい場合は相手が何を言っているのかほとんど聞き取れません．

「今日は編集で福原さんに会えるよ！」

に少しでもノイズが入って

「今日は編集でひげさんに会えるよ！」

などと聞こえたら（なわけないか・・・）大変なことです．でも

「今日（ガー）編集で（ジー）福原（ブー）んに会える（ビー)」

だったら何のためらいもなく喜んで編集に行けますね．つまり，**ノイズがあろうとなかろうと，正しく情報が伝達されることが大切で，少々ノイズが入っていても受け手の側で情報が再現できれば，伝わる情報はノイズのない完璧な通信回線と全く同じ**なのです．

これをもう少し突き詰めて考えると，ノイズの程度（伝達情報の曖昧度）というものが通信にとっては本質

的です．そしてまさにこれが「情報」の本質であり，バイオインフォマティクスにおいて「重要な部分（情報）」と言っている意味なのです．

前置きが非常に長くなりましたが，今回は Shannon あるいは Wiener といった巨人を中心とした情報理論を見てみます．そして意外にも[*1]情報理論と非常に相性がよかった生態学との関連を中心に話を展開してみます．

2 情報理論

2.1 情報の測度における 3 つの要請

情報理論では，流れてくる情報がどのような配列になるかを考えます．つまり，情報の期待値として以下 3 点の要請をみたす関数 H を求めたいということがあります．

1. H は p_i の連続関数
2. $p_i = \frac{1}{n}$ のとき，（このとき極大）H は n の単調増加関数になる．
3. ある選択が 2 つの連続な選択に分割可能である場合，全体の H は 2 つの選択に関する H_1 と H_2 の重み付き和で表される．

最後の要請が分かりにくいのですが，同じ情報の表し方が複数ある場合を考えています．「今日」と「昨日の明日」は同じ情報をもっている必要がある，というのが 3. の言っている意味です．

2. の場合を考えてみましょう．このとき $H(1/n, 1/n, \cdots, 1/n) = A(n)$ と表されるとします．等確率で s^m 個の事象が起こる場合を分解してみると，3. の要請より，$1/s$ の確率で s 個の事象が起こる過程を m 回繰り返せば同じ事になります．よって $A(s^m) = mA(s)$ となります．$t \neq s$ なる t についても，$A(t^n) = nA(t)$ としておきます．

$s^m \leq t^n < s^{m+1}$ をみたすように m を選ぶと，$A(n)$ の単調増加性より，$A(s^m) \leq A(t^n) \leq A(s^{m+1})$ すなわち

$$mA(s) \leq nA(t) \leq (m+1)A(s)$$

が成り立ちます．$s^m \leq t^n < s^{m+1}$ の各辺の対数をとると $\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n}$ より，十分大きい n に対して

$$\left| \frac{m}{n} - \frac{\log t}{\log s} \right| < \varepsilon/2$$

また $\frac{m}{n} \leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n}$ からは

$$\left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| < \varepsilon/2$$

が分かります．三角不等式：

$$\left| \frac{m}{n} - \frac{\log t}{\log s} \right| + \left| \frac{A(t)}{A(s)} - \frac{m}{n} \right| > \left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right|$$

を使うと，結局十分大きい n に対して

$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| < \varepsilon$$

[*1] 情報というものが世界の中核をなしていることを考えれば当然でもありますが．

となりますが，このとき $A(t) = -K \log t$ と表すことができます．

さて，さらに等確率 $p_i = \frac{n_i}{n} = \frac{n_i}{\sum n_i}$ で n 個の事象が生じるとします．（つまり $n_i = n_j$ です．）このとき，n 個の事象から 1 つ選択するという手続きを，

1. n 個のなかから n_i 個の集合をまず選ぶ（確率 p_i で）
2. n_i 個のなかから 1 個を選ぶ

という手順に分解します．すると $A(t) = K \log t$ を用いて[*2]

$$K \log n = H(p_1, p_2, \cdots, p_n) + K \sum p_i \log n_i$$

という式ができるので，（右辺第 2 項の p_i は重み）変形すると

$$\begin{aligned} H(p_1, p_2, \cdots, p_n) &= K \left(K \log n - \sum p_i \log n_i \right) \\ &= K \left(\sum p_i \log n - \sum p_i \log n_i \right) \\ &= -K \sum p_i \log \frac{n_i}{n} = -K \sum p_i \log p_i \end{aligned}$$

となります．ただし途中で $\sum p_i = 1$ を用いています．

これが情報のエントロピーとよばれるもので，ある事象の結果についての不確定度を表しています．名前の通り，これは熱力学でいうところのエントロピーに対応している概念でもあります．

係数 K は対象の単位によって適当に決めるものなので，以下では $K = 1$ としておきます．

3 生態学への応用

3.1 多様度

「生態系が多様である」とはどういうことでしょうか？個体数が多いとかバイオマスが・・・とかではなく，理論的に扱いやすい尺度は Shannon-Wiener 関数：

$$H = -\sum_{i=1}^{s} p_i \log_2 p_i$$

で与えられます．生物種の多様度 H は生物種数 s と，対象となっている生態系内で i 番目の種が占める割合 p_i[*3]で計算できるというわけです．ある s が与えられたときに多様度 H が最大となるのは $p_i = \dfrac{1}{s}$ のときという要請があったので，

$$H_{max} = -\sum_{i=1}^{s} \frac{1}{s} \log_2 \frac{1}{s} = \log_2 s$$

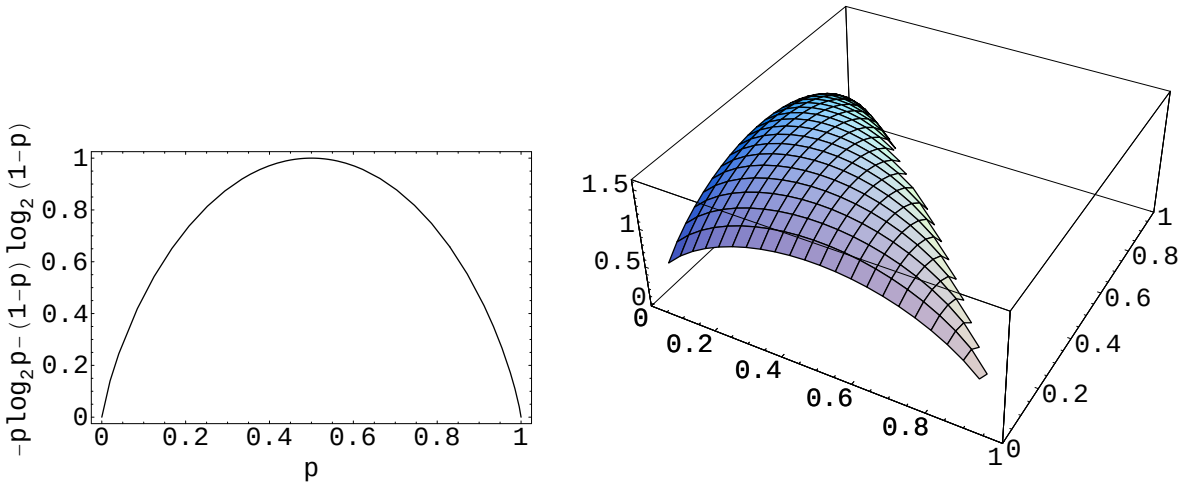
が H の最大値になります．これは直感的にも明らかで，ある生物種に偏りがあるような生態系は多様度が下がるという結果になることが分かります．

これを使うと情報理論での相対エントロピーに対応する均衡性指数 H/H_{max} が計算でき，任意の生態系に共通なパラメーターがひとつ得られたことになります．[*4]

[*2] K は $A(t)$ が単調増加になるように適当に選ぶ必要があります．

[*3] 適当に選んだ個体が「i 種」である確率と考えたらいいでしょう．

[*4] もちろん多様性を比較するのは H ですが．



(a) 2 種共存　　(b) 3 種共存

図 1　多様度のグラフ

3.2　多様度を視覚化する

2 種共存の場合の多様度：$-p\log_2 p-(1-p)\log_2(1-p)$ と 3 種共存の場合の多様度：$-p_1\log_2 p_1-p_2\log_2 p_2-(1-p_1-p_2)\log_2(1-p_1-p_2)$ をグラフに書くと**図 1** のようになります．$p=\frac{1}{2}$ や $p_1=p_2=\frac{1}{3}$ で多様度が最も高くなることが視覚的にもよく分かります．そして割合 p_i に偏りが生じると一気に多様度が落ち込んでしまうという微妙なバランス関係が見てとれます．

3.3　現実的問題

何種類かある多様性の指数のうち，もっとも使われているのは種数です．これはとにかく野外で観測しやすい量であるということが主な理由ですが，上で導入した H との違いは明白です．種数がある程度多様性を反映しているとはいえ，H のほうがより多くの情報を含んでいることは言うまでもないでしょう．

3.4　多様性の素晴らしさ

有限の環境収容力の上で H_{max} に近い状態で平衡に達し，安定している生態系にはそれだけで美しさを感じます．たとえその構成員がネズミ，ゴキブリ，コウモリ，モグラ ...etc であってもこれは変わりません．

しばしば生物をやってる人の中には，「種数が多ければいい」とか，「希少種がいればいい」といった考え方をする人がいます．もちろんこういう考え方も大事なのですが，理論としての美しさ，あるいは生物というものの全体を眺められるという点などで H がより大きいことには及ばない気がします．H には上記 2 点の意味も含まれていますしね．

4　生命の保全

4.1　ちょっと明るい未来を想像してみる

さて先ほども述べたように，多様性の指数としては種数がよく使われていて，特定の種については「北アメリカ大陸」とか「ユーラシア大陸」といったスケールで調査がなされています．しかし H での調査というのはちょっと大変です．

というわけで想像上で調査してみます．任意の生物から，その種名を個体ごとに受信できるような人工衛星が将来的に完成したとしましょう．すると地球上の全生物個体から，その位置と種名が得られます．種名から，その生態がある程度の広がりをもって分かるので，その生物種の分布が適当な解像度で計算できます．全生物種について重ね合わせをとり，あるスケールでの H を計算すれば，全地球状の多様度マップができるということになるでしょう．そして地球上が H の値で埋め尽くされるはずです．等高線を書くと視覚的に分かりやすくなるかもしれません．

このマップを見て私たちは何を感じるでしょうか？普段生活していて私たちが感じ取っている「多様度」というのはせいぜい地球の（ヒトの）人口分布くらいなものです．しかし，必ずしも京都や大阪市内の生物学的多様度が高いとは限りません．視点を変えるだけで，自然に対する意識が変わってくるはずです．

最近話題の外来生物についても考えてみてください．彼らを捕獲し，何らかの処置をすることにためらいもあるかもしれませんが，彼らの進出によって崩れていく生態系のバランス・多様度という観点からみると，考えも変わってくるはずです．先ほどのグラフでも分かるとおり，多様性が失われた後で多様度を上げるということは非常に困難です．自然に与える人間の影響力は甚大なのです．

4.2　最後に

以上で終わりにしたいと思いますが，バイオインフォマティクスや生態学に限らず，情報理論というものはあらゆる学問の中核をなしているような気がします．自分の興味がある学問分野で，情報というものが与えてくれる描像に思いを巡らせるのも楽しいものです．

誤植，意見，苦情等は，次のアドレスまで．

ichino@mathscphys2004.mbox.media.kyoto-u.ac.jp

参考文献

[1] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.
http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

[2] 寺本英, 岩波講座現代物理学の基礎（第 2 版）『生命の物理』第 II 部, 岩波書店, 1978.

[3] Charles J. Krebs, ECOLOGY, Benjamin Cummings, 2001.