行政院國家科學委員會專題研究計畫 成果報告

保守非編碼序列的數學模型解釋及功能判別方式
研究成果報告(精簡版)

計 畫 類 別 : 個別型
計 畫 編 號 : NSC 98-2621-B-006-004-
執 行 期 間 : 98 年 08 月 01 日至 99 年 07 月 31 日
執 行 單 位 : 國立成功大學生物多樣性研究所

計 畫 主 持 人 : 吳煥麟

計 畫 參 與 人 員 : 博士班研究生-兼任助理人員：曾聖智
博士班研究生-兼任助理人員：賴建志

處 理 方 式 : 本計畫可公開查詢

中華民國 99 年 07 月 20 日
1. Introduction
Cross-species genome comparison is a fundamental method for identifying biologically essential elements. Using this method, a considerable number of conserved non-genic sequences (CNGs) have been identified in vertebrates (Margulies et al., 2003; Bejerano et al., 2004 and Dermitzakis et al., 2005) and also insects (Glazov et al., 2005). Some CNGs are even more conserved than coding sequences (Bejerano et al., 2004; Dermitzakis et al., 2003), which might indicate that at least some CNGs are functionally important. Experimental attempts to test the hypothesis that CNGs are previously unidentified genes showed that this is unlikely (Dermitzakis et al., 2002; Mural et al., 2002; Frazer et al., 2001 and Waterston et al., 2002). Several other potential CNG functions have been suggested, including a role as regulatory regions (Duret et al., 1993; Hardison et al., 1997 and Hardison, 2000) and matrix attachment regions (Glazko et al., 2003). There is evidence, moreover, that CNGs are important in vertebrate development (Woolfe et al., 2005). Deletion of 1817 kb and 983 kb gene deserts (gene-poor regions greater than 500 kb) containing a total of 1243 CNGs in mice, however, resulted in no detectable phenotypic changes and only small changes in the expression of genes that are adjacent to the deletions (Nobrega et al., 2004). Recently An experiment done by Ahituv et al. showed that removal of four non-coding ultraconserved elements from the mouse genomes made no recognizable differences (Ahituv et al., 2007).

2. Statistical evidence
We hypothesise that most CNGs have never been under selective pressure and that substitution rate is the only dynamic force on them. This hypothesis is supported by the results presented in Figures 2.1 and 2.2 (using previously published data, Table 1): average CNG length is essentially constant across all human chromosomes except the Y chromosome whereas average conserved exon length is much more variable (Figure
and average CNG density (chromosome length divided by number of CNGs) varies little across chromosomes 1-18 and 20 (the range of values obtained is 7382 to 9663, where 7382 indicates that on average there is a CNG every 7382 bases), whereas the density of conserved exons is more variable (range 7471 to 25308) across chromosomes 1-18 and 20 (Figure 2.2). As an interesting aside, the results of these simple calculations show that most of the smaller chromosomes, including the Y chromosome, tend to have distinct properties (Figures 2.1 and 2.2). The numbers of conserved exons and CNGs on the Y chromosome, for example, are 602 and 993 (the other chromosomes have between 1837 and 20713 exons each and between 2159 and 32427 CNGs each), so there may be a bias in the data for the Y chromosome due to its small size. It is possible that the numbers of conserved exons and CNGs on the Y chromosome are low at least partly because there is only one copy of the Y chromosome.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of Conserved exons</th>
<th>Average conserved exon length</th>
<th>Number of CNGs</th>
<th>Average CNG length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20713</td>
<td>292.22</td>
<td>27890</td>
<td>151.45</td>
</tr>
<tr>
<td>2</td>
<td>18029</td>
<td>264.50</td>
<td>32427</td>
<td>160.67</td>
</tr>
<tr>
<td>3</td>
<td>13241</td>
<td>281.05</td>
<td>26257</td>
<td>154.33</td>
</tr>
<tr>
<td>4</td>
<td>9090</td>
<td>289.36</td>
<td>22184</td>
<td>154.62</td>
</tr>
<tr>
<td>5</td>
<td>10675</td>
<td>297.70</td>
<td>24521</td>
<td>155.13</td>
</tr>
<tr>
<td>6</td>
<td>10519</td>
<td>303.66</td>
<td>19813</td>
<td>151.25</td>
</tr>
<tr>
<td>7</td>
<td>10616</td>
<td>284.14</td>
<td>18895</td>
<td>153.64</td>
</tr>
<tr>
<td>8</td>
<td>7742</td>
<td>297.22</td>
<td>16300</td>
<td>153.46</td>
</tr>
<tr>
<td>9</td>
<td>8893</td>
<td>293.69</td>
<td>14679</td>
<td>153.78</td>
</tr>
<tr>
<td>10</td>
<td>9456</td>
<td>280.01</td>
<td>16123</td>
<td>153.08</td>
</tr>
<tr>
<td>11</td>
<td>12086</td>
<td>298.70</td>
<td>16471</td>
<td>151.84</td>
</tr>
<tr>
<td>12</td>
<td>10890</td>
<td>282.46</td>
<td>13541</td>
<td>148.81</td>
</tr>
<tr>
<td>13</td>
<td>4466</td>
<td>295.70</td>
<td>11696</td>
<td>157.87</td>
</tr>
<tr>
<td>14</td>
<td>7015</td>
<td>301.80</td>
<td>11859</td>
<td>153.48</td>
</tr>
<tr>
<td>15</td>
<td>8261</td>
<td>282.79</td>
<td>10620</td>
<td>153.66</td>
</tr>
<tr>
<td>16</td>
<td>9032</td>
<td>283.00</td>
<td>9670</td>
<td>159.58</td>
</tr>
<tr>
<td>17</td>
<td>11557</td>
<td>285.36</td>
<td>9127</td>
<td>147.63</td>
</tr>
<tr>
<td>18</td>
<td>3921</td>
<td>274.32</td>
<td>9331</td>
<td>156.12</td>
</tr>
<tr>
<td>19</td>
<td>8540</td>
<td>334.70</td>
<td>2264</td>
<td>156.11</td>
</tr>
<tr>
<td>20</td>
<td>4904</td>
<td>314.30</td>
<td>7084</td>
<td>151.13</td>
</tr>
<tr>
<td>21</td>
<td>1837</td>
<td>326.60</td>
<td>2978</td>
<td>150.02</td>
</tr>
<tr>
<td>22</td>
<td>3894</td>
<td>331.30</td>
<td>2159</td>
<td>142.65</td>
</tr>
<tr>
<td>X</td>
<td>8244</td>
<td>298.63</td>
<td>23295</td>
<td>153.20</td>
</tr>
<tr>
<td>Y</td>
<td>602</td>
<td>284.00</td>
<td>993</td>
<td>220.50</td>
</tr>
</tbody>
</table>

Table 1. Numbers of conserved exons and CNGs for each chromosomes
(using previously published data; Dermitzakis et al., 2005).

Figure 2.1. Comparison of the average lengths of conserved exons and CNGs for each human chromosome.

Figure 2.2. Comparison of the densities of conserved exons and CNGs for each human chromosome.

The densities were obtained by dividing the number of bases in a particular chromosome by the number of exons/CNGs on that chromosome.
Superficial examination of the data in Figures 2.1 and 2.2 indicates that exon evolution has been somewhat chromosome-dependent whereas CNGs have evolved largely independently of which chromosome they are on. Therefore we consider further whether CNGs and exons have been shaped by the same dynamic force. Hence we calculate the correlation between the average lengths of conserved exons and CNGs. In order to try to answer this question, we checked the correlation between average exon length and average CNG length. (The data used are the same as those used for Figure 2.1).

The Pearson correlation coefficient is written:

\[ r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y} \]

where \( \bar{x} \) and \( \bar{y} \) are the sample means of \( x_i \) (the average length of conserved exons) and \( y_i \) (the average length of conserved CNGs), \( s_x \) and \( s_y \) are the sample standard deviations of \( x_i \) and \( y_i \) and the sum is from \( i = 1 \) to \( n \) (\( n=24 \) with chromosome Y and \( n=23 \) without chromosome Y).

Using this formula and data from Table 1, the correlation coefficient is \(-0.13\) without the Y chromosome and \(-0.25\) with the Y chromosome included. The average lengths of conserved exons and CNGs for each human chromosome therefore have a negative and low correlation.

These statistical analyses, suggests that even if CNGs have been shaped by selection, then selection pressure has been exerted in a different way on CNGs than on exons. Logically, however, it is more likely that most CNGs have not been shaped by selection. Therefore we have built a simple mathematical model using Ramsey Theory to test whether long-term conservation of numerous stretches of non-coding DNA could have occurred under no selective pressure.

3. Ramsey theory model
Humans are predisposed to finding patterns but some patterns can arise by chance from randomness. Ramsey Theory is used to explain those patterns that arise from randomness; it concerns the conditions under which patterns must appear and, indeed, implies that complete disorder is impossible. Problems in Ramsey Theory typically
ask a question of the form: how many elements of some structure must there be to guarantee that a particular property will hold? Perhaps the most famous case explained by Ramsey theory is The Bible Code (Figure 3.1). It has been claimed that the Hebrew text of Genesis contains intentional coincidences of words or phrases that appear as letters with equal spacing (Drosnin, 1997 and 2003). However McKay et al. subsequently disproved this claim by proving that the same code can be found in the first 78,064 letters (the same length as Genesis) of the Hebrew translation of War and Peace (Figure 3.2; Bar-Hillelet al., 1998; McKay et al., 1999). There are about 3 billion letters in our ‘life book’, the human genome. Is this sufficiently large to find patterns such as CNGs in the same way that patterns can be found in the Bible or any other sufficiently large text? We use Ramsey Theory and some previous results concerning substitution rates to indicate that yes, it is likely that many CNGs have arisen by chance.

Figure 3.1. One of examples mentioned in Dorsnin’s book, The Bible Code.
Figure 3.2. McKay made a joke on Drosnin. He found a message in Moby Dick that Drosnin would be killed.

Some studies (Smith et al., 2002; Hardison et al., 2003) suggest that substitution rates vary deterministically across the non-coding non-repetitive regions of the human genome. Properties of substitution rates from the non-coding part of the human genome were identified using human-chimpanzee pair-wise comparisons over 4.7 Mb of sequence. These properties are (1) regional variation, with significant differences between alignments of tens of kilobases on the same chromosome; (2) local similarity, with adjacent one kb blocks within alignments tending to have similar substitution rates; (3) repeatable variation, with the substitution rate variation down the human and chimpanzee lineages being positively correlated; (4) no evidence to indicate that the observed substitution rate variation is owing to selection. We think that regional variation gives rise to characteristics such as CNGs, with regions with slower substitution rates being the source of most CNGs.

The probability of finding even one ultra-conserved element, let alone 481 or even 327,000, has been calculated as vanishingly small (less than $10^{-22}$) (Bejerano et al., 2004); a mean of the Bernoulli variable for substitution rate from a 1 Mb window was used to determine this probability. We argue that the local substitution rates in this analysis were overestimated because an average value was used, ignoring regional variations, and that the implications of Ramsey Theory were ignored. Assuming the four results of Smith and co-workers (Smith et al., 2002), we can show that most of the 327,000 CNGs actually have a high probability of occurring. Employing Ramsey Theory and assuming neutral substitution rates, we calculate (see Section 4) that a region of a 300 million year old ancestor sequence of only 12970 bases in length is required for humans and mice to retain a CNG of at least 200 bp with 100% identity and 95% identity to chicken. Our analysis indicates, therefore, that many CNGs have arisen by chance and are non-functional. Identification of functional CNGs may be facilitated by comparison of genomes from evolutionarily distant species such as human and fugu (Woolfe et al., 2005; Goode et al., 2005), since the numbers of CNGs shared between species decreases with evolutionary distance.
4. Assessment of likelihood that CNGs have arisen by chance
Suppose a genomic region with substitution rate $r$ and length $\ell$ (in nucleotides) existed 300,000,000 years ago (the approximate time of a common ancestor of birds and mammals - Figure 4.1). From the time that the human and mouse lineages split (80,000,000 years ago), what is the value of the length $\ell$ required, according to Ramsey theory, in order that human and mouse inherit an identical 200 bp segment and that chicken inherits this segment with 95% identity to human and mouse?

**Figure 4.1.** Illustration of the hypothetical situation used to assess the likelihood that a CNG can arise by chance. A 300 million year (myr) old genomic region is represented by the long yellow line at the top of the diagram. The human and mouse lineages split approximately 80,000,000 years ago and the avian and mammalian lineages lineage split approximately 300,000,000 years ago. A 200 bp segment that has been inherited by human and mouse with 100% identity and by chicken with 95% identity to human and mouse is represented by the short yellow line towards the bottom of the diagram.

Firstly, the substitution rate, defined as the number of substitutions per site per year, is calculated as follows:

$$K = 2Tr,$$

where $K$ is the number of substitutions between two homologous sequences (species), $T$ is the time of divergence between the two sequences (species), and $r$ is defined as
the number of substitutions per site per year. The substitution rate is therefore given by

$$2 \times 3 \times 10^8 \times r = 0.05$$

where $3 \times 10^8$ is the number of years and 0.05 is the total change due to substitutions (for retention of 95% identity with chicken).

This gives a substitution rate in substitutions per site per year of

$$r = 8.3333 \times 10^{-11}$$

With this substitution rate estimate, the chance of a nucleotide change since the time when the co-ancestor of the human and mouse lineages lived (80,000,000 years ago) is:

$$8.3333 \times 10^{-11} \text{ substitutions per site per year} \times 8 \times 10^7 \text{ years} = 6.6666 \times 10^{-3}.$$

The probability of a nucleotide base remaining the same is then $(1 - 6.6666 \times 10^{-3}) = 0.9933$, and therefore the probability of a region of 200 nucleotides remaining absolutely conserved is:

$$(0.9933)^{200} \times (0.9933)^{200} = 0.0679.$$  

As the size of the genomic region ($N$) increases, the probability of finding an ultra conserved region of 200 nucleotides increases. If we pick the first base to 200th base in this genomic region, the second base to 201st base and so on, then there is $(200 + 1)$ events in total (Figure 4.2). Hence it will be expected there are 0.0679 $(200 + 1)$ CNGs in region with length N if these ultra conserved segments are disjoint one another.

![Figure 4.2](Image)

Figure 4.2 The conserved genomic region (indicated in yellow) may be
selected from the first base to 200th base, the second base to 201st base, and so on.

However theses segments are jointed. Actually we should subtract the numbers of intersection of any 2 of segments; add the numbers of intersection of any 3 of segments; subtract the numbers of intersection of any 4 of segments; add the numbers of intersection of any 5 of segments, and so on (Figure 4.3).

**Figure 4.3.** Venn diagram of three sets. The number of CNGs is

\[ |A_j| \quad |A_j \cap A_i| + \quad |A_j \cap A_i \cap A_r| \]

Now let us calculate each term of  

\[ |A_j| = (0.9933)^{400} \quad (200 + 1) \ldots (1) \]

But  

\[ |A_j \cap A_i| = (0.9933)^{402} \quad (201 + 1) \quad (0.9933)^{404} \quad (202 + 1) \]

\[ (0.9933)^{406} \quad (203 + 1) \quad \ldots \quad C(j,0) \quad (0.9933)^{2k} \quad (k + 1) \ldots (2) \]

\[ |A_j \cap A_i \cap A_r| = (0.9933)^{404} \quad (202 + 1) + 2 \quad (0.9933)^{406} \quad (203 + 1) \]

\[ + \ldots + C(j,1) \quad (0.9933)^{2k} \quad (k + 1) \ldots (3) \]
\[ |A_i \cap A_i \cap A_i \cap A_k| = (0.9933)^{406} \quad \binom{203 + 1}{3} \quad (0.9933)^{408} \quad (203 + 1) \quad \ldots \quad C(j,2) \quad (0.9933)^{2k} \quad (k + 1) \quad \ldots \quad (4) \]

\[ |A_i \cap A_i \cap A_i \cap A_k \cap A_p| = (0.9933)^{408} \quad \binom{204 + 1}{4} \quad (0.9933)^{410} \quad (205 + 1) \quad \ldots \quad \binom{c(j,3)}{2k} \quad (k + 1) \quad \ldots \quad (5) \]

where \( C(j,r) = \frac{j!}{r!(j-r)!} \), and \( j! = j \times (j-1) \times (j-2) \times \ldots \times 2 \times 1 \).

Checking their coefficients of corresponding terms, such as (\(-1+2-1\)) \((0.9933)^{406} \quad (203 + 1)\), these coefficients will be cancelled out. Hence \((1)+(2)+(3)+\ldots+(n)\) will get \((0.9933)^{400} \quad (200 + 1)\) \((0.9933)^{402} \quad (201 + 1)\) and other terms from those terms with higher power, such as those terms from \((n-1)\) and \((n)\).

Therefore if we take a much bigger number \(n\) \((n=600\) as an example), then those terms with higher power will vanish.

And \((1)+(2)+(3)+\ldots+(n) = (0.9933)^{400} \quad (200 + 1)\) \((0.9933)^{402} \quad (201 + 1)\).

Hence we get the inequality and calculate it to obtain the low bound \(N\), the length we expect to have at least one 200 bp CNG.

\[(0.9933)^{400} \quad (200 + 1) \quad (0.9933)^{402} \quad (201 + 1) > 1.\]

Hence \(N=1236\) is obtained.

This result is perhaps counter-intuitive, but human intuition focuses on a specific outcome, the 200 bp conserved segment, ignoring the fact that we are calculating the probability of this outcome from a huge sample space.

We note that the substitution rate for CNG sequences estimated above \((8.3333 \times 10^{-11}\) substitutions per site per year) is 25-44 times slower than a recent estimate of the average rate of substitution in mammalian DNA of \(2.1-3.7 \times 10^{-9}\) substitutions per year. Bejerano et al. calculate that CNG elements are changing at a rate that is roughly 20 times slower than the average for the genome. Taking these figures into account, we estimate that an upper value of the substitution rate in CNGs might be something like

---

This text contains mathematical expressions and calculations relevant to the analysis of probabilities and substitution rates in CNG sequences.
\[ r = 8.3333 \times 10^{-11} \quad (44/20) = 1.8333 \times 10^{-10} \]

With this substitution rate, the chance of a substitution is:

\[ 1.8333 \times 10^{-10} \times 8 \times 10^{-7} = 1.4666 \times 10^{-2} \]

The probability of a nucleotide base remaining the same is 0.9853 and, therefore, the probability of 200 bp that are absolutely conserved between human and mouse is

\[ (0.9853)^{200} \times (0.9853)^{200} = 2.6754 \times 10^{-3} \]

From the relationship

\[ (0.9853)^{400} \times (200 + 1) - (0.9853)^{402} \times (201 + 1) > 1 \]

let N=600 if N<600,

We find that even with the faster substitution rate, a region of the ancestor sequence with length greater than or equal to only 12970 bp is needed for occurrence of a 200 bp conserved element with 100% identity between human and mouse and 95% identity between human/mouse and chicken.

5. References


無研發成果推廣資料
國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估
   ■達成目標
   □未達成目標（請說明，以 100 字為限）
       □實驗失敗
       □因故實驗中斷
       □其他原因
   說明：

2. 研究成果在學術期刊發表或申請專利等情形：
   論文：□已發表 ■未發表之文稿 □撰寫中 □無
   專利：□已獲得 □申請中 ■無
   技轉：□已技轉 □洽談中 ■無
   其他：（以 100 字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

在生物資訊的方法中，常利用序列保守程度來預測序列是否具有功能。那些比蛋白質編碼區還要保守的 CNGs 顯示它們可能具有重要的功能。然而二個最近的研究顯示，剔除 3,000,000 個鹽基序列（其中有 1,243 個 CNGs）和剔除四個超級保守序列的小鼠與正常小鼠並無顯著差異，似乎指出有些 CNGs 並不具有重大功能。本研究計畫首先建立數學模型去解釋這一矛盾現象，提出 CNGs 的產生如「聖經密碼」一樣，有可能是隨機過程的必然結果。因為 CNGs 不參與蛋白質和非編碼 RNA 的合成，研究其功能的方法主要是靠剔除法（knockout）：然而 CNGs 的数目是編碼基因的十幾倍，在操作剔除法前若能對具有功能的 CNGs 進行篩選，將大大減少時間與人力上的消耗。根據已建立的數學模型，本研究計畫提出一個不須藉由保守程度來判斷 CNGs 是否具有功能的方法。