CG-SSR:跨物種比對平臺之 SSR 資料庫

CG-SSR: an online comparative genomics database for SSR discovery

蕭孟昌，陳建銘，蕭子焚，高華霖，白敦文*
國立台灣海洋大學資訊工程學系
基隆市北寧路 2 號，中華民國台灣
*E-mail: twp@mail.ntou.edu.tw

摘要

簡單重複序列(simple sequence repeat, SSR)指的是一段以 2 ~ 6 個鹼基對做為重複單位的 DNA 片段，對於基因的調控網路扮演了極為重要的角色，並且廣泛的應用在基因體的相關研究。CG-SSR(comparative genomics database for SSR discovery)資料庫提供使用者查詢各個模型生物的 SSR 序列，並提供該 SSR 與基因序列間精確的相對及絕對位置資訊。然而電腦預測 SSR 序列的結果仍舊是存有高度的偽陽性，因此我們採用了比較基因體學的方法針對判讀出的 SSR 序列進行篩選，使用者可以自由選定物種和標的物種進行比對，座落在跨物種間保留區塊的 SSR 才予以保留，此方法明顯的提升了預測結果的精確度。CG-SSR 資料庫提供友善的使用介面與服務，每一筆 SSR 資料都有詳盡且實用的資料連結，充分的符合從事基因體研究者的需求。

關鍵詞：簡單重複序列、比較基因體、保留區域、跨物種比對、資料庫

Abstract

Simple sequence repeats (SSRs), also referred to as variable number of tandem repeats or micro-satellites, are valuable genetic markers which play a crucial role in genome mapping and various genetic studies. In this study, we have set up a database which facilitates the search for SSRs and provide absolute and relative location information of corresponding genes. However, performing *in silico* analysis of biological data sometimes attempts to result in high false positive rates. In order to promote the specificity of discovering important SSRs from our proposed system, we take advantage of evolutionarily conserved segments among sequences from various species. Users are able to choose specific species as targets to filter out SSRs which are not located in conserved regions. Screening processes narrow down candidate SSRs and improve the performance of specificity of characteristics. In this database, there are eleven representative species collected for comparative genomics analysis. Taking the comparison between zebrasifh and fugu as an example, 38,773 SSRs from zebrafish genome were found

located in conserved regions in which 9.35% SSRs are found in protein-coding regions, 0.30 % in 5'UTR, 1.27% in 3'UTR, 50.63% in intron, and 38.45% in intergenic region. Each SSR is precisely allocated and annotated in this database for further applications.

Keywords：Simple Sequence Repeat (SSR), comparative genomics, conserved region, cross-species comparison, database

一、前言

隨著人類基因體的解碼後，基因體相關的研究也隨之蓬勃發展，其中 SSR(simple sequence repeat) 又稱為微衛星序列(microsatellites)，是一段以 2 ~ 6 個鹼基對做為重複單位的 DNA 片段，在真核生物和原核生物的染色體內大量的出現[10, 12, 20]，因其具有含量豐富、多對偶基因、高度多型性、再現性高和容易以 PCR 偵測等優點，成為廣泛應用於分子育種及遺傳多樣性評估的分子標記，近年的研究更發現 SSR 對於基因調控扮演了極為重要的角色，影響了基因活性、DNA 複製、細胞週期、基因重組等功能，甚至掌控了生物演化與環境調適的功能。目前已廣泛的應用於物種育種、遺傳疾病檢測、物種鑑定、族群研究、基因圖譜及親子鑑定等方面之研究[14]。

近年來有進一步的研究指出，位於蛋白質編碼區(protein coding region)的 SSR，其數量的增減會造成基因的框架位移突變(frame shift mutation)，進而影響了基因功能性的表現[16]。SSR 位於 5'端非轉譯區域(5'UTR)的部分影響了基因的轉錄和轉譯作用，進一步影響了基因的表現[18]。位於 3'端非轉譯區域(3'UTR)的 SSR 造成了轉錄位移(transcription slippage)和 mRNA 數量的增加，其結果可能會造成 mRNA 堆積在細胞核內從而阻斷了截切作用(splicing)和其他的細胞功能[4, 19]。而位於內顯子(intron)內的 SSR 可能會影響基因轉錄，mRAN splicing 和 mRNA 運送至細胞質的功能[6, 8, 15]。三倍體的 SSR 座落於 UTR 或 intron 區間會造成 heterochromatin-mediated-like 基因的不表現[2, 3, 7, 9]。這一些現象直接都受到了 SSR 序列的調控，並且明顯的影響了基因的表現。

最新的研究指出，SSR 甚至可以視為生物體調節環境變異的調節器。生物體利用 SSR 來控制部分的基因表現與否，進而達到環境適應和物種演化的功能[13]。因為 SSR 對於基因的調節扮演了如此重要的角色，也因此越來越多的研究者投入相關領域的研究。本資料庫提供了各個物種的 SSR 序列和每一個 SSR 與基因體對應的相關資訊，並且註明其相對應的 primer 資訊做為研究者進行 PCR 實驗的使用。然而電腦僅能針對 SSR 序列的重複性等相關特性進行分析，許多電腦判讀的 SSR 未必在生物體內扮演重要的功能性，如此的分析造就了高度偽陽性的結果。為了要增加預測的精確度，本資料庫提供跨物種的比對模式以供使用者進行篩選。因為就比較基因體學的角度而言，基因體內具有功能性的區塊承受較大的演化壓力，也因此相對的較不容易突變。將各個物種的序列進行比對的結果發現，各物種間序列一致性較高的區塊往往是決定其基因功能性的區塊。因此倘若使用者所找出的 SSR 座落於跨物種比對後的保留區塊(conserved region)內，其 SSR 扮演調節基因功能性的機會必然相對的提高。使用者初步的 SSR

搜尋結果經過比較基因體的過濾後，所得到的結果將保留較具有生物功能性的 SSR 序列，此過濾功能亦明顯的提升了電腦判讀的準確度。

二、研究方法

CG-SSR 資料庫的資料產生流程如圖一所示，使用的 SSR 序列和 SSR primer 等資訊皆取自於澳洲的 SSR Primer 和 SSR Taxonomy 資料庫(http://bioinformatics.pbcbasc.latrobe.edu.au/cgi-bin/ssr_taxonomy_browser.cgi)[11]，但是該資料庫所提供的 SSR 有重複計算的冗餘資料存在，且沒有提供正確的相對位置資訊進行驗證，使用者無法從資料庫中迅速判斷任一個 SSR 在基因體中的正確位置資訊，所以本系統將以該 SSR 之資料為基礎，去除了重複計算的 SSR 片段後，再透過 Ensembl 資料庫(http://www.ensembl.org/index.html)所提供之基因座標，對於 SSR 的序列加以註記其位於基因體的相關位置，包含了位於基因的上游序列(upstream)、下游序列(downstream)、5'UTR、3'UTR、protein-coding region、intron 以及基因間序列(intergenic region)等座標。因為基因體內存在了選擇性截切(alternative splicing)以及 RNA 編輯(RNA editing)等作用，使得每一筆基因序列可能對應到多筆的轉錄序列，為了避免資料重複記錄的問題產生，每一筆基因序列保留了所對應最長的轉錄序列。此外，為了提升 SSR 序列判讀結果的精確度，本資料庫採用比較基因體學的方式針對初步判讀的結果進行篩選。而跨物種比對的共同保留區域資訊取自於 UCSC Genome 資料庫(http://genome.ucsc.edu/cgi-bin/hgGateway)。此資料庫提供了跨物種比對基因體內保留區域之座標資訊，此資料和初步篩選出的 SSR 結果進行比對，再將座落於保留區域內的 SSR 予以保留，如此經過不同物種 SSR 的完整蒐集，去冗餘性分析，正確位置標記及跨物種保留區之比較分析，使用者所得到篩選過後的 SSR 序列將是高度具有生物功能性的 SSR 序列。

三、操作介面設計

本資料庫系統經上述之流程產生之後，為了讓使用者可以依不同需求進行檢索，查詢介面設計特別與生物學者進行討論並訂定查詢選項之規格。如圖二所示，使用者可以自由選取欲搜尋的物種，特定的染色體或基因座標，SSR 片段的型態與長度，CG-SSR 資料庫將會針對使用者所設的條件參數進行查詢並顯示符合條件的所有搜尋結果，包含 SSR 序列的基本型態、重複長度、基因座標位置和所有 SSR 對應的正反向 primer(此由 SSR Taxonomy 資料庫所提供之資訊)。另一個重要的跨物種比對查詢的介面如圖三所示，使用者可以選定特定的物種和原有物種進行比較基因體篩選，使用者亦可以設定 SSR 的長度和座落在保留區域內的重疊比例等參數對初步的結果進行過濾，此查詢功能將提供位於不同物種保留區域之內的 SSR 序列及其完整的相關資訊。在此選定斑馬魚做為標的物種，河豚做為比較基因體的篩選物種，共同保留區域的門坎比例設為 80%，SSR 序列長度最短為 20 個鹼基，以此條件所得到的搜尋結果如圖四所示，每一筆 SSR 的長度、位置、序列型態等資訊都有清楚的標示與適當的連結。

四、討論

本資料庫系統將收錄 11 種具代表性物種的完整 SSR 資訊，經事先的分析、註記及交叉比對後，使用者可以快速的查詢所有資訊。在此以斑馬魚為例進行說明，由表一可以得知，以斑馬魚為主體和其他十個物種進行跨物種的比對，斑馬魚和同屬魚類的稻田魚、刺魚、河豚和虎河豚等魚種，其保留區域佔斑馬魚基因體的比例，較兩棲類的非洲爪蟾和哺乳類的大鼠、小鼠、負子鼠、人以及鳥類的雞等物種來的高，如此的數據充分符合我們對於物種親緣關係的認知。而表二所探討的是斑馬魚進行跨物種比對後，其 SSR 座落於保留區域(conserved region)內的比例。如圖表所示，斑馬魚的 SSR 座落於同屬魚類的物種保留區域(conserved region)的比例仍舊較兩棲類、鳥類及哺乳類物種來的高，除了大鼠的比例略高於河豚之外。整體的分析結果依舊符合物種演化的親源性。

表三顯示了符合比較基因體篩選的 SSR 位於基因體的分佈情形，多數的 SSR 是座落於 intron 區域，其次是基因與基因之間的 intergenic region(僅設定在基因前後各 15K 的範圍內而言)，再其次為 protein-coding region，座落於 5'UTR 和 3'UTR 的比例相對較低。

以斑馬魚為例，我們提供了十種模式物種供使用者進行比較基因體的篩選。挑選河豚與斑馬魚進行跨物種比對的結果，原本 715,225 筆的 SSR 片段經過比較基因體的篩選後，共有 38,773 筆斑馬魚的 SSR 序列座落在保留區塊內，其中 9.35%的 SSR 序列座落於 protein-coding region，0.30 % 座落於 5'UTR, 1.27%座落於 3'UTR，50.63% 位於 intron, 以及 38.45% 位於 intergenic region。此結果亦符合對於現有 SSR 性質的認知。

除了實現資料庫系統提供個別物種及跨物種比對的 SSR 查詢系統外，本論文亦提出數個具有實驗佐證的實際範例進行說明。在表四所展示的五個案例，使用者能透過 CG-SSR 資料庫找出 SSR 序列，並進一步的通過了比較基因體學的篩選，所得到的結果皆獲得了實驗證實。以人類的基因體為例，使用老鼠做為比較基因體的篩選物種。第一列與第二列以(CAG)為基本型態重複發生的 SSR 序列，分別座落在 HD 及 DRPLA 基因上，前者造成了杭丁頓氏舞蹈症，後者則造成相當罕見的體染色體顯性遺傳的神經退化性疾病。第三列資料是以(CGC)為基本型態重複發生的 SSR 序列，座落在 PABP2 基因上，造成了眼咽肌肉失養症。第四列則是以(A)為基本型態重複發生的 SSR 序列，座落在 MBD4 基因。造成了 MMR 基因的不活化，進一步的引發癌症的產生。第五列則是以(A)為基本型態重複發生的 SSR 序列，座落在 BLM 基因，此 SSR 具有抑制癌症的功能。以上述範例可以得知本資料庫所提供的 SSR 序列對於基因體的調控確實扮演了決定性的角色，CG-SSR 資料庫不但提供了友善的操作介面與詳盡的基因體資訊，且較其他 SSR 資料庫更多出比較基因體篩選的功能，能更精確的預判出扮演基因調控決定性的 SSR，本資料庫對於基因體的相關研究將是一大助益。

誌謝：本論文之完成感謝國立台灣海洋大學水產生物科技頂尖研究中心之計畫經費贊助。

五、參考文獻

[1]. B.Brais, J.P.Bouchard, Y.G.Xie, D.L.Rochefort, N.Chretien, F.M.Tome, R.G.Lafreniere, J.M.Rommens, E.Uyama, O.Nohira, S.Blumen,

A.D.Korczyn, P.Heutink, J.Mathieu, A.Duranceau, F.Codere, M.Fardeau, and G.A.Rouleau. Nat. Genet. **18,** 164 (1998).

[2]. C.J.Cummings and H.Y.Zoghbi. Annu. Rev. Genomics Hum. Genet. **1,** 281 (2000).

[3]. B.M.Davis, M.E.McCurrach, K.L.Taneja, R.H.Singer, and D.E.Housman. Proc. Natl. Acad. Sci. U. S. A **94,** 7388 (1997).

[4]. L.Deng and S.Shuman. Biochemistry **36,** 15892 (1997).

[5]. A.Duval, S.Rolland, E.Tubacher, H.Bui, G.Thomas, and R.Hamelin. Cancer Res. **60,** 3872 (2000).

[6]. Y.Ejima, L.Yang, and M.S.Sasaki. Int. J. Cancer **86,** 262 (2000).

[7]. E.Fabre, B.Dujon, and G.F.Richard. Nucleic Acids Res. **30,** 3540 (2002).

[8]. N.Gabellini. Eur. J. Biochem. **268,** 1076 (2001).

[9]. A.M.Gacy, G.Goellner, N.Juranic, S.Macura, and C.T.McMurray. Cell **81,** 533 (1995).

[10]. D.W.Hood, M.E.Deadman, M.P.Jennings, M.Bisercic, R.D.Fleischmann, J.C.Venter, and E.R.Moxon. Proc. Natl. Acad. Sci. U. S. A **93,** 11121 (1996).

[11]. E.Jewell, A.Robinson, D.Savage, T.Erwin, C.G.Love, G.A.C.Lim, X.Li, J.Batley, G.C.Spangenberg, and D.Edwards. Nucleic Acids Research **34,** W656 (2006).

[12]. H.Karaoglu, C.M.Lee, and W.Meyer. Mol. Biol. Evol. **22,** 639 (2005).

[13]. Y.Kashi and D.G.King. Trends in Genetics **22,** 253 (2006).

[14]. Y.C.Li, A.B.Korol, T.Fahima, and E.Nevo. Molecular Biology and Evolution **21,** 991 (2004).

[15]. C.L.Liquori, K.Ricker, M.L.Moseley, J.F.Jacobsen, W.Kress, S.L.Naylor, J.W.Day, and L.P.Ranum. Science **293,** 864 (2001).

[16]. D.Metzgar, J.Bytof, and C.Wills. Genome Res. **10,** 72 (2000).

[17]. K.Nakamura, S.Y.Jeong, T.Uchihara, M.Anno, K.Nagashima, T.Nagashima, S.Ikeda, S.Tsuji, and I.Kanazawa. Hum. Mol. Genet. **10,** 1441 (2001).

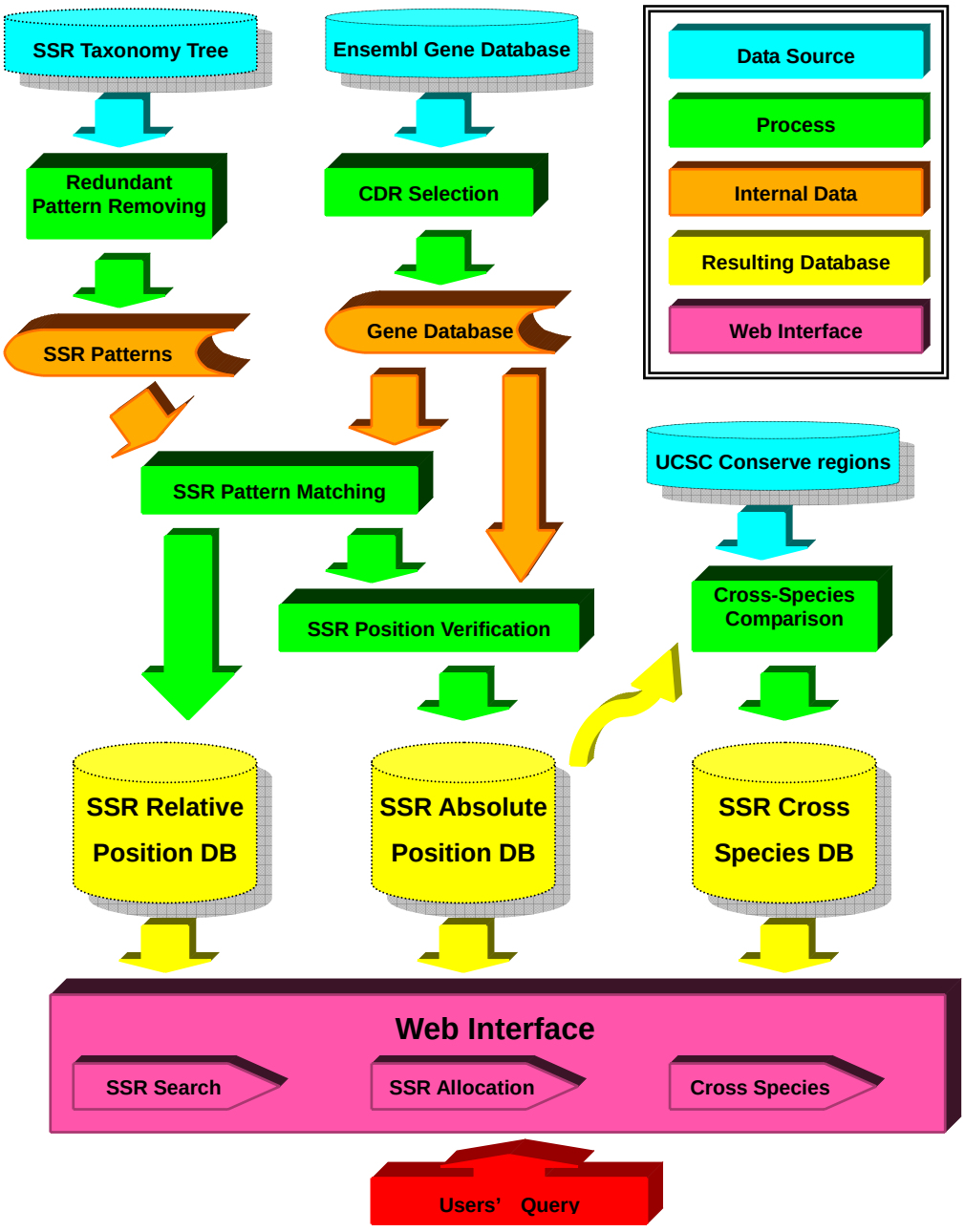
[18]. G.Raca, E.Y.Siyanova, C.T.McMurray, and S.M.Mirkin. Nucleic Acids Res. **28,** 3943 (2000).

[19]. N.Suraweera, B.Iacopetta, A.Duval, A.Compoint, E.Tubacher, and R.Hamelin. Oncogene **20,** 7472 (2001).

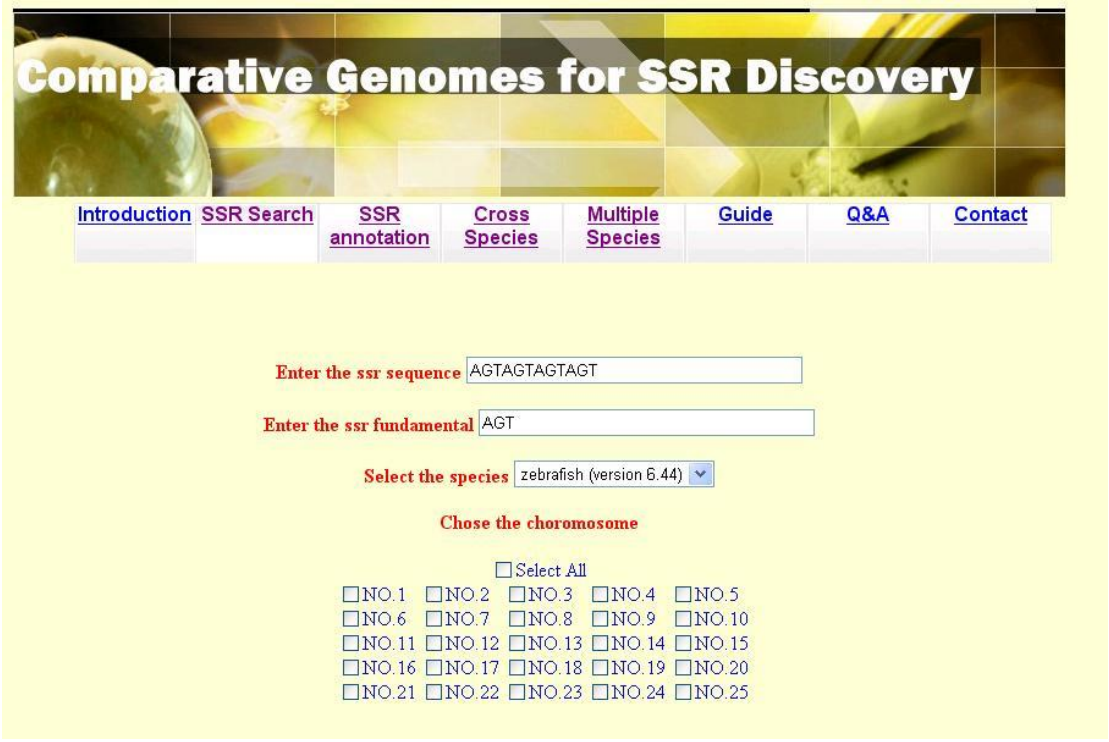
[20]. G.Toth, Z.Gaspari, and J.Jurka. Genome Res. **10,** 967 (2000).

[21]. M.Yamada, S.Tsuji, and H.Takahashi. Ann. Neurol. **52,** 498 (2002).

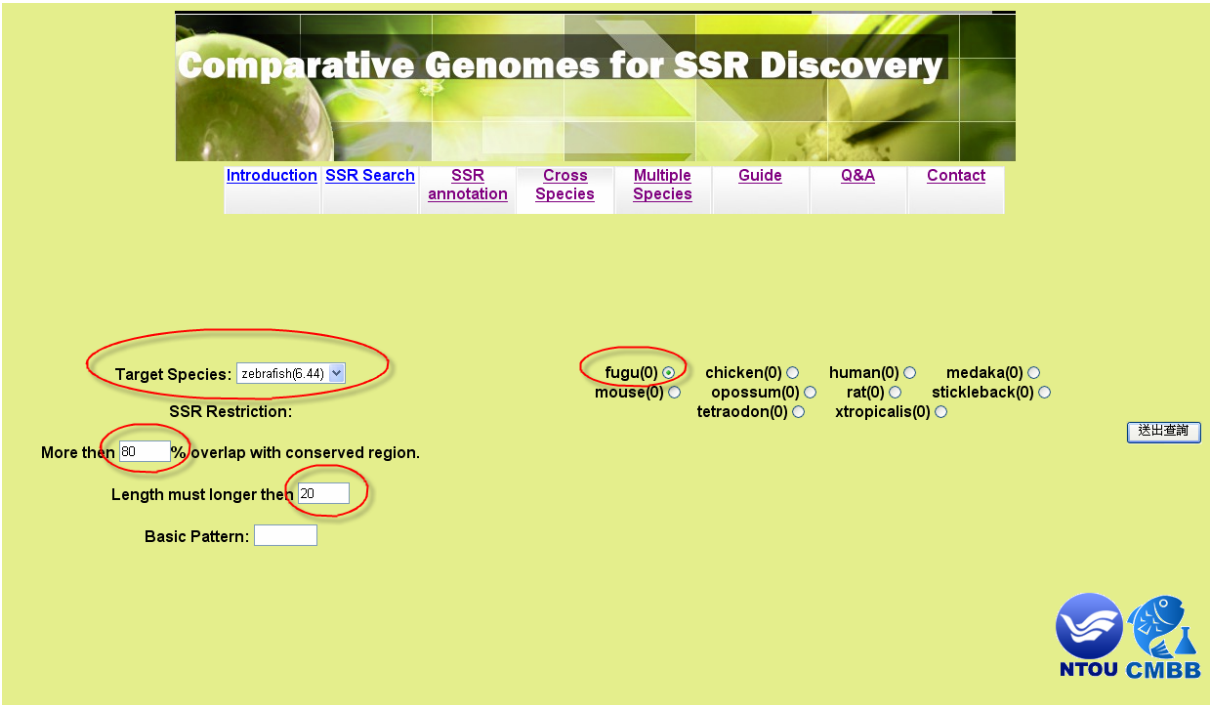
[22]. H.Y.Zoghbi and H.T.Orr. Annu. Rev. Neurosci. **23,** 217 (2000).



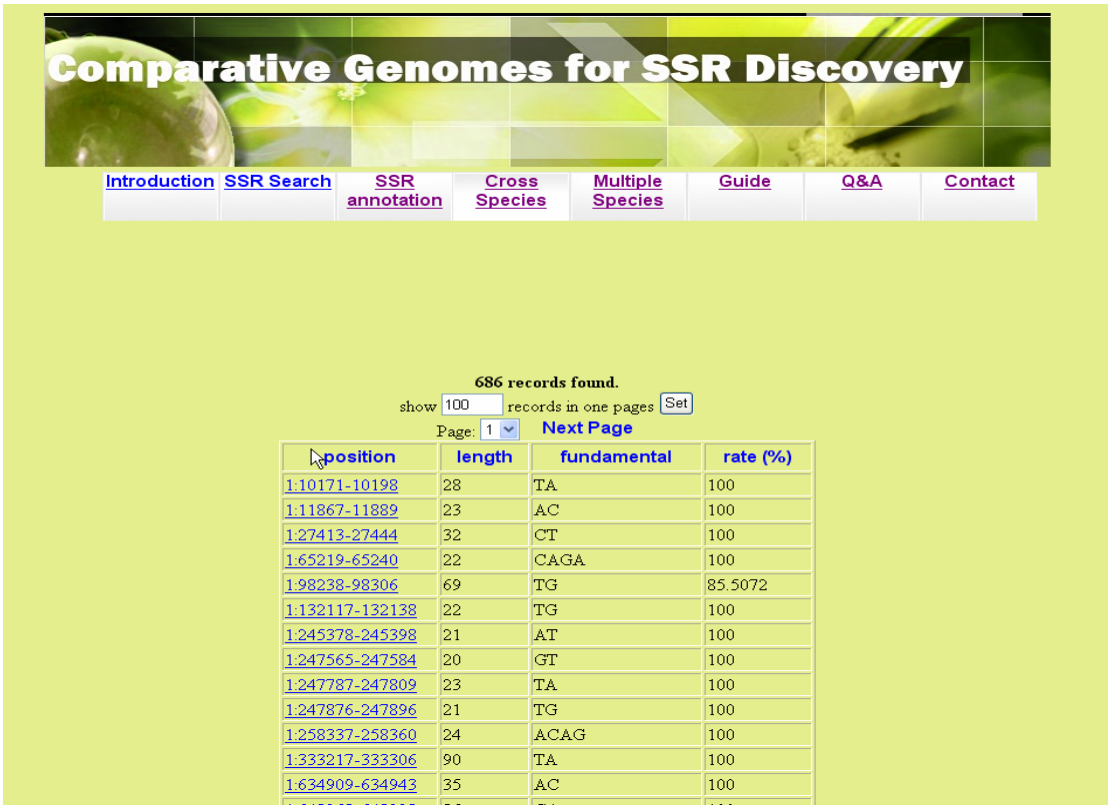
圖一、CG-SSR 資料庫之資料產生流程圖。



圖二、CG-SSR 資料庫的使用者介面，使用者可以根據物種、染色體位置、SSR 序列型態等選項或是針對特定基因等參數進行搜尋特定的 SSR 序列。

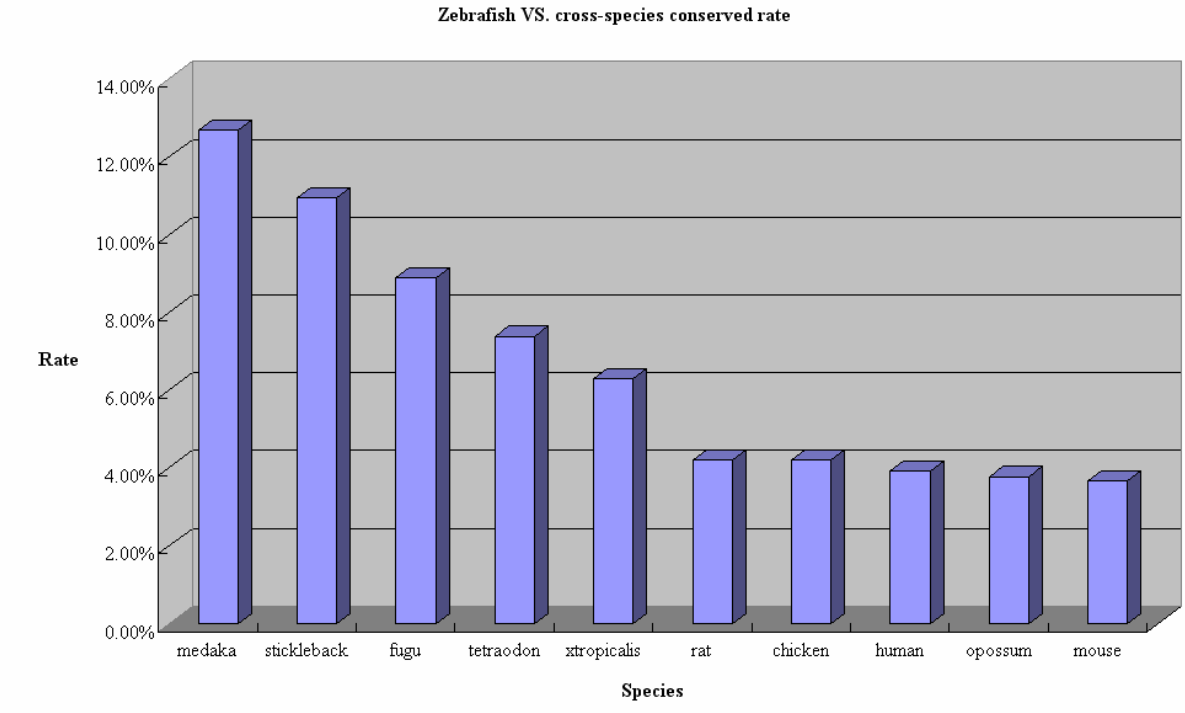


圖三、使用者可以選定標的物種進行跨物種比對，透過保留區的比對讓初步搜尋出的 SSR 結果進行過濾，提高預測結果的精確度。

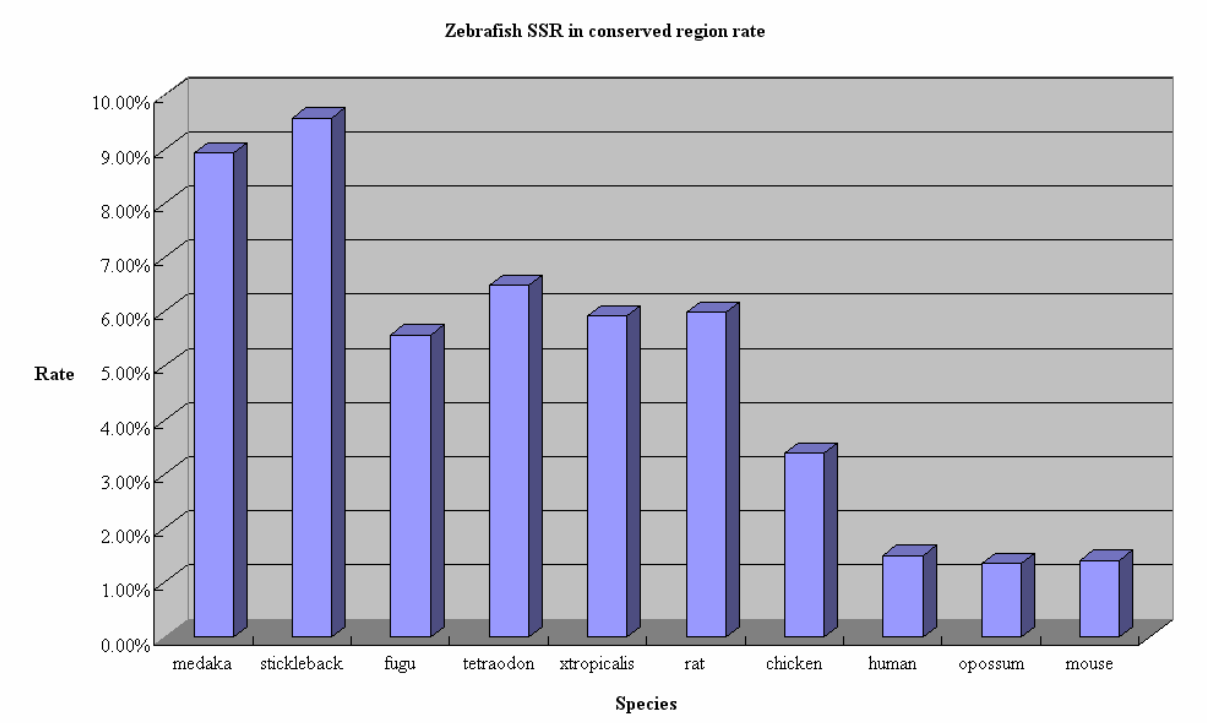


圖四、CG-SSR 資料庫根據使用者所選定的參數，顯示符合條件的結果。

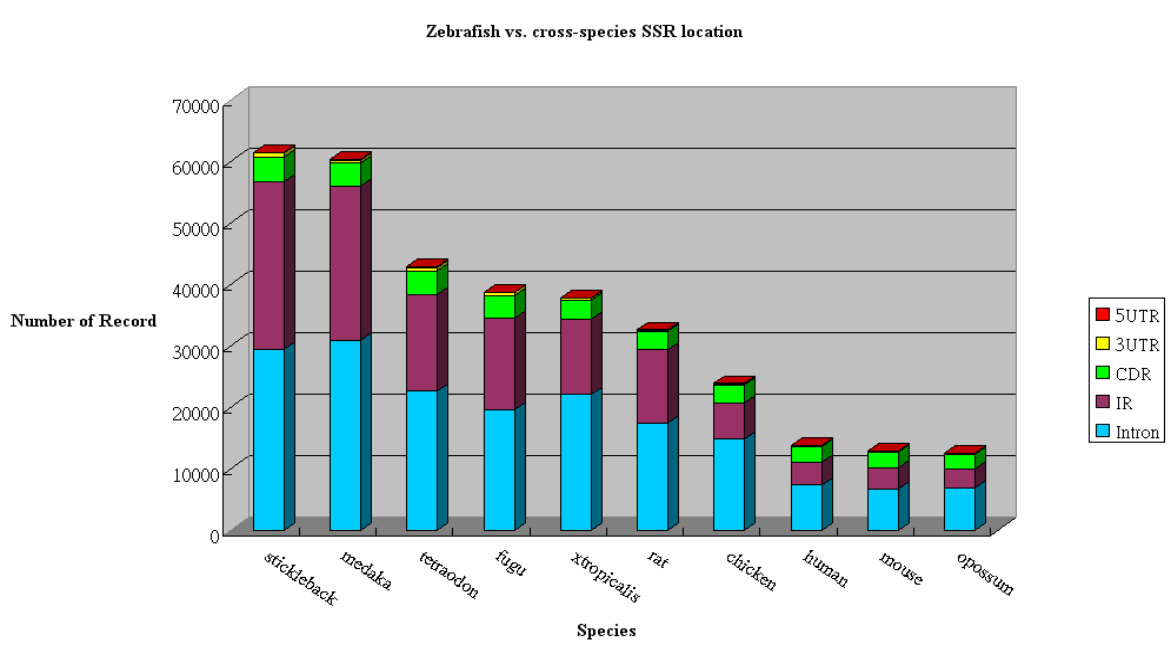
表一、班馬魚和十個物種進行跨物種比對，跨物種保留區域佔斑馬魚基因體的比例。



表二、班馬魚和十個物種進行跨物種比對，屬於班馬魚的 SSR 座落於跨物種保留區域的比例。



表三、班馬魚和十個物種進行跨物種比對，班馬魚的 SSR 座落於跨物種保留區域內於基因體的分佈情形。



表四、通過跨物種比對門坎所篩選出的 SSR，且已有實驗證實具有功能的 SSR 序列及相關論文。

Target Species	Comparative Species	Motif	Gene	SSR Function	Reference
Human	Mouse	(CAG)n	HD	Expansion causes Huntington's disease (HD)	[22,7]
Human	Mouse	(CAG)n	DRPLA	Causes dentatorubropallidoluysian Atrophy (DRPLA)	[17,8]
Human	Mouse	(CGC)n	Poly(A)-binding protein 2 (PABP2)	Oculopharyngeal muscular dystrophy	[1,7]
Human	Mouse	(A)n	MBD4	Frameshift caused by (A)n size changes inactivates MMR genes and causes human cancers	[21,3]
Human	Mouse	(A)n	BLM	Tumor-suppressive function	[5,9]