しかし，これらタンパク質の結晶化方法は，主に構造解析を専門とする研究者によって開発されてきたため，X線結晶構造解析のルーツである，有機低分子化合物の構造解析をテーマとしてきた専門家らが用いた結晶化法が応用され，例えば，結晶成長中においては，溶液を静置することが重要で，装置の振動等を極力排除することが重要と考えられてきた．このため，結晶品質に悪影響を与えるとされる，重力による浮力対流を排除するため，宇宙空間での結晶育成が行われている．このほか一般的な結晶品質の向上のための方策として，結晶化サンプルの純度や鮮度，沈殿剤濃度，pH，添加剤，蒸気拡散の速度，温度，結晶化ドロップの容量，結晶化溶液の混合方法，結晶化方法，結晶化用の容器，試薬の製造会社などが考慮されてきたが，試行錯誤する必要があった．それでも，これら様々な結晶化の条件を試す必要があることから，サンプル量の少量化，ハイスループット化が進められたが，膜タンパク質を中心に結晶化が非常に困難なものも多く，インシリコ創薬のボトルネックとなっていた．

そこで，大阪大学工学研究科を中心に，フェムト秒レーザー照射による結晶核発生技術，溶液かく拌による高品質結晶育成技術を駆使した，従来とは全く異なるタンパク質結晶育成技術を開発し，その技術を基に2005年に大学発ベンチャーとして㈱創晶を設立し，公的なサンプルについては共同研究として結晶解析する創晶プロジェクトも設けている．

2-1. レーザー照射による核発生制御技術の開発

大阪大学の佐々木研究室では，以前から電界センサーやテラヘルツ波発生特性が優れている有機非線形光学結晶4-dimethylamino-*N*-methyl-4 stilbazolium tosylate（DAST）の高品質化技術の研究を実施している．[1-3] 結晶そのものを有機材料として利用するには，大型化，高品質化が必要で，過飽和度の高い状態で核を発生させ，結晶を立てて成長させるなど工夫が施されたが，自然核発生を完全には制御できず，結晶サイズの不均一性など材料としは不向きな点が課題として残り，低過飽和溶液からの結晶核の発生が古くより検討された．しかし，核発生は，一般に物理的には2次の相転位だという以外にメカニズムが未解明で，何か刺激が必要だという以外に解決策がなかった．同研究室でレーザーアブレーションによるAlN薄膜やBCN薄膜の研究が行わていたのが幸いし，[4-7] 容器中の溶液に刺激を与えるのにはレーザーがよいという発想に至っている．ただし，1996年のPhysical Review Letters（PRL）に報告されているように，高過飽和尿素溶液にNd：YAGレーザーを照射すると電界効果で結晶が析出するという内容の論文[8]が掲載されており，核発生にレーザー照射が有効だという，厳密な第一発見者ではないものの，レーザー照射を低過飽和の溶液で行い，結晶成長に利用するという初めての試みがなされた．その結果，通常は自然核発生が起こらない低過飽和度溶液からでも，Nd：YAGレーザー照射により核発生を誘起できることが判明した．

2-2. 溶液かく拌による高品質結晶育成：$CsLiB_6O_{10}$の実験　紫外レーザー光は，機械材料・構造体等のマクロ加工，電子産業分野での超微細加工，半導体リソグラフィー用光源，目の屈折矯正手術（LASIK）などの医用等，多くの分野にその応用が期待されているが，従来の紫外レーザー光源である稀ガスハライド系のエキシマーレーザーでは多くの課題が残されていた．これに対して森らは，紫外線レーザー開発を目的として1993年に組成が$CsLiB_6O_{10}$(CLBO）の新しい非線形光学材料の発見に成功した．[9-11] しかし，高出力紫外光が発生すると，自身の発する紫外レーザーにより出力の低下やレーザー損傷が問題となった．可能な限りレーザー損傷耐性の高いCLBO結晶の育成技術の開発が必要となり，最終的には結晶育成条件の最適化が重要と考え，融液や溶液の制御を行い，結晶育成の本質に迫ることが試みられた．

1998年には，種々の検討の結果，CLBO溶液全体をかく拌しながら育成することで高品質化が試みられ，出力が通常の方法よりも最大で2.5倍向上することが分かった．その後，溶液かく拌条件の最適化を徹底的に行ってCLBO結晶の高品質化・均一化がより進み，この高レーザー耐力CLBO結晶を用いて，三菱電機㈱と共同で196 Wの高繰り返しグリーンレーザー光（波長：532 nm，繰り返し10 kHz）から42 Wという世界最高出力のNd：YAGレーザーの第4高調波（266 nm光）発生に成功することができた．[12] この結果は，高出力全固体紫外光源の実用化という観点から非常に意義深いものであるとともに，溶液かく拌という手法が高品質結晶

化に効果的だということを示したという点でも重要であったと思われる．最近では，三菱電機と共同で10 W を超える Nd：YAG レーザーの 5 倍高調波（213 nm）光発生，ニコンと共同で 1547 nm 光の第 8 高調波（193 nm）発生にも成功している．

2-3. 新規結晶化技術のタンパク質の結晶化への応用 タンパク質の結晶化が他の材料に比べ難しい理由の 1 つに，準安定領域が極めて大きく，過飽和度をかなり高くしないと自然核発生が起こらず，運よく核を得ても高過飽和溶液中での結晶成長は結晶品質に悪く，大量析出による多結晶化などの問題が生じる．したがって，より低過飽和の溶液における強制的な結晶核の発生と，低過飽和の溶液のままの持続的な結晶育成が重要である．森らは，DAST の結晶化で試みたレーザー照射の技術や，CLBO で実績を積んだ溶液かく拌による高品質結晶育成技術を，低過飽和溶液でのタンパク質の結晶化に適用し，従来とは全く異なるタンパク質結晶育成技術を開発した．

高野，安達，森らは，まず最初にリゾチームを使って溶液かく拌技術の開発に着手し，従来法よりも早い成長速度で大きな結晶ができることを見い出した．一方，レーザー照射による結晶核発生に関しては，Nd：YAG レーザーの出力を上げても，可視光や紫外レーザー光の照射を試みても駄目であったが，工学研究科応用物理学専攻の増原研究室の協力を得て，フェムト秒レーザーを照射する実験を試みたところ，照射回数に応じて発生した結晶核の数が増減することが判明した．さらに，グルコースイソメラーゼを使った実験で，従来法では結晶核が発生しない低過飽和度溶液でも，フェムト秒レーザーの照射により結晶核を強制的に発生させることに成功した．当時筆者らは，トリパノソーマ由来プロスタグランジン $F_{2\alpha}$ 合成酵素の結晶化に苦労していたが，フェムト秒レーザー照射により，数ヵ月掛かった結晶化がわずか 2–3 日で結晶核の発生に成功し，フェムト秒レーザーの照射が結晶核の発生に効果があることが実証できた．[13]

アステラス製薬㈱との共同研究で，human triosephosphate isomerase（TIM）の結晶化実験にこれらの技術を適用したところ，従来法で 2.8 Å しか得られなかったが，溶液かく拌法のみの適用で，1.2 Å 分解能まで X 線回折が起こり（Fig. 1），最終的に 1.4 Å の構造解析に成功した．[14]

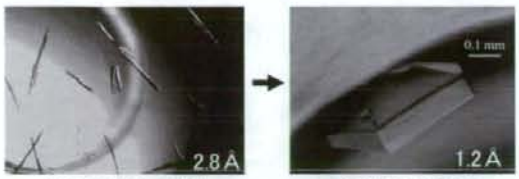

Fig. 1. Human Triosephosphate Isomerase Crystals Obtained with the Conventional (Left) and Stirring Methods (Right)

2-4. 膜タンパク質の結晶化等への応用 2003 年には，阪大産研の村上が創晶プロジェクトに参加し，2002 年に *Nature* の表紙を飾った大腸菌由来多剤排出トランスポーター膜タンパク質 AcrB[15] の結晶化にフェムト秒レーザーの照射技術が試された．通常は結晶核の発生が起こらない，低過飽和度溶液に照射したところ，強制的な結晶核の発生に成功し，結晶からの回折分解能が 3.5 Å から 2.3 Å に，さらにレーザー照射結晶核発生後，溶液かく拌しながら育成すると 2.1 Å 分解能まで飛躍的に向上した（Fig. 2）．[16]

また，膜タンパク質 SecDF の高品質結晶育成（従来法の 5.6 Å 分解能から 3.74 Å 分解能に向上）[17] の結晶化においても，レーザー核発生と溶液かく拌の 2 つの技術の有効性が実証された．

一般に，膜タンパク質の結晶化を行うためには，高純度に精製し，さらに高濃度に濃縮する必要がある．水溶性タンパク質と異なり，精製の段階から界面活性剤等を使用して可溶化する必要がある．膜に埋もれた状態から可溶化されて生体中とは異なる環境下に置かれるため，大きな不安定化要因となる．少しでもサンプルの劣化を防ぐために，結晶化では高濃度の沈殿化剤が用いられ，結晶育成のための時間を長く取れず，これが膜タンパク質の結晶の品質に悪影響を与えていると思われる．

レーザー照射による核発生までの時間を節約し，結晶育成中でのかく拌による高品質化に時間を掛けることができる点で，レーザーかく拌とかく拌の組み合わせは，まさに膜タンパク質に最適な結晶化技術である可能性が高い．

一方，tRNA 修飾酵素である MnmA と tRNAGlu の複合体（Fig. 3）結晶の高品質化（従来法では 4～5 Å であった分解能が 3.1 Å 分解能に向上）な

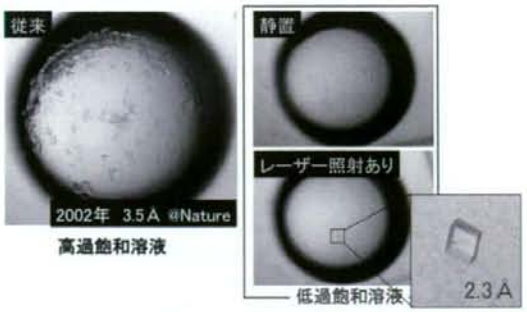

Fig. 2. Typical Crystallization Results for AcrB Crystals Obtained from a Highly Supersaturated Solution (Left)
Conventional growth (without laser irradiation, top in the right panel) and laser-irradiated growth (bottom in the right panel) in a low-range supersaturated solution.

Fig. 3. Crystal Structure of an RNA Sulfuration Enzyme Mnma Thiouridylase-tRNA Complex

ど，高品質の結晶が得られ難いサンプルに対しても有効であった.[18,19] 今後益々結晶化が困難なサンプルについて依頼が増えると予想され，さらなる高効率な結晶化技術の開発が望まれている.

3. 通常の *In Silico* での探索手法とその問題点

インシリコ創薬では，計算機上で，標的タンパク質の基質結合部位に，化合物ライブラリーに含まれる多数の薬物分子を順次結合させ，複合体構造と結合自由エネルギー（ΔG）を予測し，ドッキングスコア（以下スコアと呼ぶ）のよい化合物をヒット化合物候補として採択する．ドッキング計算における ΔG の予測誤差は約 3 kcal/mol あり，ヒット化合物の示す ΔG 値とほぼ同程度であるため，インシリコスクリーニングのヒット化合物予測精度は，ランダムスクリーニングよりはましとは言える．しかし，それでもヒット率はいまだに低い．その理由は，複合体構造の予測が困難であるからである．たとえ，タンパク質—化合物ドッキング計算を，長時間の分子動力学計算（MD）を利用して行っても，誤った予測構造からの ΔG の計算は実測値との差が大きく，ヒット率は低下してしまう．複合体構造予測については，実用に供するほどの精度を持った一般的な手法は成熟していないのが現状である.

3-1. 新規なインシリコ創薬の手法の開発と創薬バリューチェインの構築 NPO 法人バイオグリッド関西の研究グループでは，タンパク質—化合物相互作用行列を作成し，これを用いた新しいインシリコ創薬の手法である Multiple target screening (MTS) 法[20] と Docking score index (DSI) 法[21] を開発し，ヒット化合物予測率を飛躍的に向上させることに成功しており，新しいインシリコでの化合物探索手法の確立が進んでいる．さらに，大阪大学客員教授である坂田らを中心に，㈱セルフリーサイエンス，創晶プロジェクト，㈱プロテインクリスタルなどのベンチャーのほか，NEC や富士通九州など大手メーカーも結集して創薬バリューチェインと呼ばれる研究グループを結成し，製薬企業との橋渡し役を担うことを目的として，公的研究機関等からの新規な標的タンパク質を用いたインシリコ創薬が可能な研究チームが編成された（Fig. 4）．まずはその特徴的なインシリコ探索の技術について簡単に述べる.

3-2. Multiple target screening (MTS) 法[20] 通常のインシリコスクリーニング手法においては，標的タンパク質に結合する化合物を，結合の強さを示すスコアの高い順に選択するが，いまだヒット率は低い．その理由の 1 つとして，標的タンパク質に対し高いスコアを示す化合物を選択すると，その化合物は他のタンパク質に対しても高いスコアを示すことが多く，かならずしも標的タンパク質に対して選択的に強い結合性を示さないことが挙げられる．これに対し，福西，中村らが開発した MTS 法では，

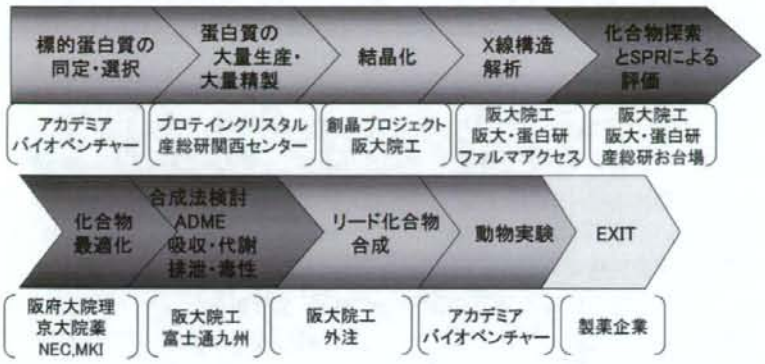

Fig. 4. The Crystal Structure of Orotidine 5′-monophosphate Decarboxylase from *Plasmodium falciparum*

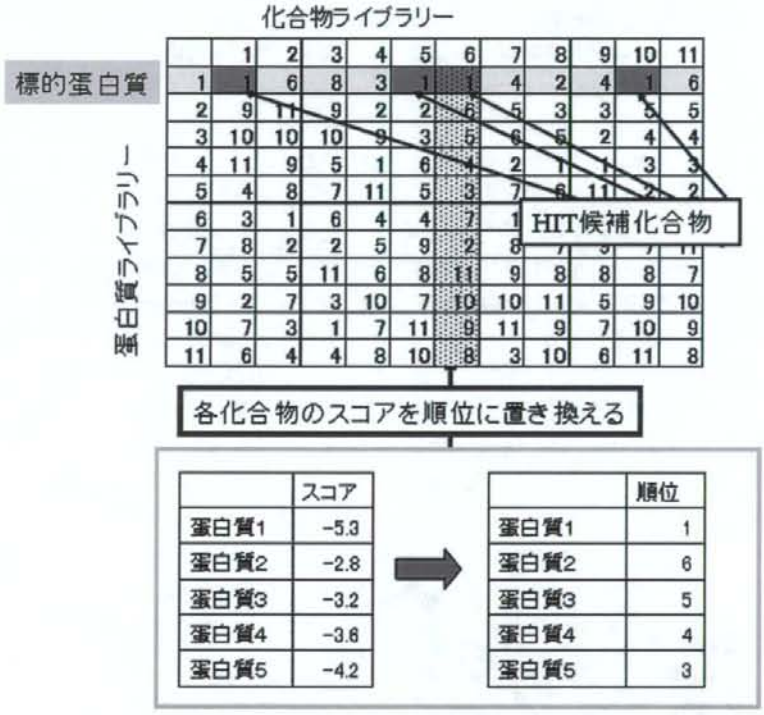

Fig. 5. The Flow Chart of Pharamaceutical Innovation Value Chain

まず，化合物ライブラリーに対して，標的タンパク質以外にも多数のタンパク質を用意し，総当たり式にドッキング計算を行って，タンパク質—化合物相互作用行列を作成し，個々の化合物が多数のタンパク質に対して結合する度合いを調べ，標的タンパク質に最も強く結合する化合物をヒット化合物の候補とする手法を考案している（Fig. 5）．詳細は，福西らの書いた総説などを参照されたいが，[20] 実際に，COX-2 や HIV プロテアーゼ 1 など 5 種類の標的に対して，MTS 法を適用し，既知の阻害剤を探索することを行った結果，計算で予測した上位 1% の化合物群の中に，ランダムスクリーニングで探索する場合と比べて，約 40 倍のエンリッチメントを得ることが示されている．

3-3. Docking score index (DSI) 法[21] DSI 法は，MTS 法と同様にタンパク質—化合物相互作用行列を用い，既知の活性化合物と類似の化合物を検索する方法として開発された．MTS 法が，ドッキング計算のため標的タンパク質の 3 次元立体構造を必要とするのに対して，DSI 法ではこれを必要とせず，MTS 法で用いたタンパク質—化合物相互作用行列を用いて，化合物の類似性検索を行う．通常，化合物の類似性は，分子量や疎水性等の分子の物理化学的性質を表す指標で示されるが，ここでは，多くのタンパク質の結合ポケットへの結合の度合いを示すスコアの集合から，統計的に類似性を抽出する．したがって，DSI 法では，標的タンパク質の立体構造の情報の代わりに阻害剤の情報が必要で，これが多いほど統計的な類似性検索の予測精度が上昇する．本方法は，G タンパク質共役型受容体（GPCR）のように，立体構造の解析が困難な膜タンパク質に対して阻害剤検索を行う場合に特に威力を発揮すると期待されている．

実際に，マクロファージ遊走阻害因子（macrophage migration inhibitory factor; MIF）等の水溶性タンパク質に 4 種類の GPCR を加えた 9 種の標的タンパク質に対して，立体構造が未知という条件の下，それぞれの標的タンパク質に対する活性化合物の情報を基に DSI 法を適用した結果，化合物ライブラリーから予測上位 1％の化合物を採択したとき，ランダムスクリーニングに比べて平均約 70 倍のエンリッチメントを得ることが示された．

3-4. 創薬バリューチェインでの実施例 創薬バリューチェインにおける最初の研究実施例は，熱帯熱マラリア原虫 *Plasmodium falciparum* 由来 Orotidine 5′-monophosphate 脱炭酸酵素の阻害剤探索で得られた．まず創晶プロジェクトにおいて結晶化を行い，筆者らの研究室において 2.6 Å 分解能での構造解析を行った（Fig. 6）．[22-23]

続いて，本酵素の構造情報を利用した MTS 法，及び基質類似の阻害剤の構造情報を利用した DSI 法を適用し，新規骨格を有した阻害剤の探索に関する研究を共同で行った．

まず，化合物ライブラリーについては，あとで購入することを考慮して，ナミキ商事㈱のカタログから提供されている化合物リストを用い，分子量等が適度な 100 万個の化合物から，MTS 法，DSI 法のそれぞれの手法を適用し，上位 5000 化合物リストアップした．共通してリストされた化合物もあったため，その和集合を取り，7622 個まで絞込んだ．続いて，NEC 社の佐久間，高田らにより，MMPB (Poisson-Boltzuman) SA (surface area) 法による再ドッキング計算を行い，上位 200 個程度までを発注した．一方，京都大学の奥野らが，7622 個の化合物の構造情報を基にして，kmeans 法や kPCA 法を使って 200 種類に分類し，富士通九州の北島らが，

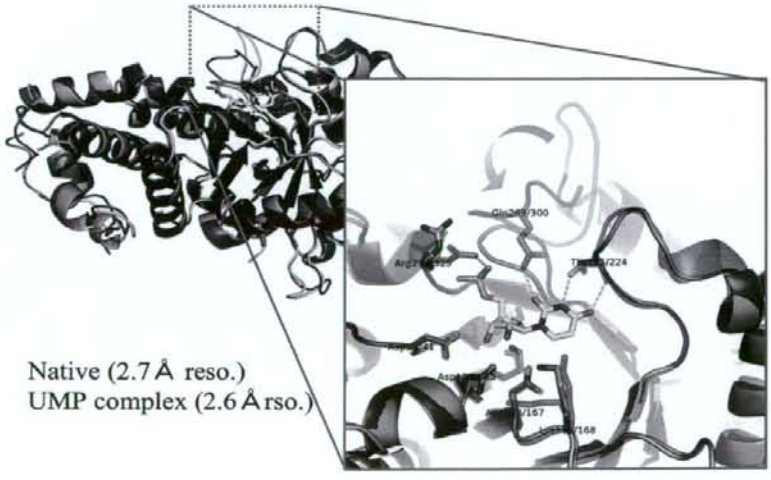

Fig. 6. The X-ray Structure Analysis of Orotidine 5′-monophosphate Decarboxylase from *Plasmodium falciparum*

その分類毎，また化合物毎に，溶解性，LogP 等々の物性情報を付加する作業を分担した．

先の，200 個発注した中で，156 個を購入し，実際に活性測定を行った．100–300 μM の濃度における 1 次スクリーニングでは，156 化合物中 33 個の化合物が酵素活性を 100% 阻害した．続いて，濃度を段階的に下げた状態での活性測定を行い，IC_{50} 値を測定することができた化合物は，156 個中 15 個存在した．10% 近い高ヒット率で阻害剤候補化合物が得られたことになる（Fig. 7）．

現在，先の 200 種類に分類した化合物群の中に，さらに強い阻害剤がないか化合物の購入と阻害率の測定を検討中である．また，15 種類得られた阻害剤候補化合物との複合体の X 線構造解析も進行中で，類似化合物の複合体構造と阻害活性の相関関係について解明し，ハイブリッド化による強力な阻害剤の開発を予定している．

なお，本手法は，別の酵素を例に，既にある製薬企業によっても実証研究が進められており，100 万化合物の中から 3229 個の候補化合物を検索し，先行して入手できた 915 個の化合物のうち，濃度 90 μM で阻害活性 50% 以上の分子が 335 個得られた．このうち，濃度依存試験を行って IC_{50} 値の測定を行ったところ，100 μM 以下を示した分子が 27 個，20 μM 以下を示したのが 8 個発見された．IC_{50} 値が 100 μM 以下を示した分子は合計で 35 個となり，やはり高ヒット率が証明されている．

謝辞　本研究で各種ソフトウェアを利用した阻害剤探索における実際の計算に当たり，創薬バリューチェインのメンバーである，日本電気㈱基礎・環境研究所の佐久間俊広・高田俊和両氏をはじめ，㈱富士通九州システムエンジニアリングの北島正人・湯田浩太郎両氏，三井情報開発㈱の福岡良忠・奥村利幸両氏らに大変お世話になり，この場をお借りして深謝致します．なお，創晶プロジェクトに関する研究は，文部科学省の大学等発ベンチャー創出支援制度をはじめ，NEDO，JST，大阪北部（彩都）地域知的クラスター創成事業「フェムト秒レーザーを用いたタンパク質結晶化と結晶加工に関する技術開発」などの支援を受けました．一方，創薬バリューチェインに関する研究では，大阪北部（彩都）地域知的クラスター創成事業「インシリコでの創薬手法の確立とその実証研究」の一環として行われたものであります．この場をお借りて感謝致します．

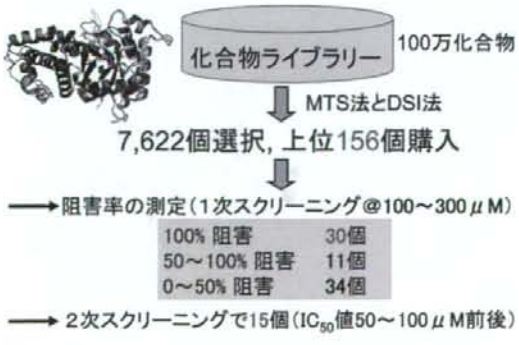

Fig. 7. Results of the *In Silico* Screening with the MTS and DSI Methods for Orotidine 5′-monophosphate Decarboxylase from Human Malaria Parasite *Plasmodium falciparum*

REFERENCES

1) Adachi H., Takahashi Y., Yabuzaki J., Mori Y., Sasaki T., *J. Cryst. Growth*, **198/199**, 568 (1999).
2) Mori Y., Takahashi Y., Iwai T., Yoshimura M., Yap Y. K., Sasaki T., *Jpn. J. Appl. Phys.*, **39**, L1006–L1008 (2000).
3) Adachi H., Nagaoka K., Tsunesada F., Yoshimura M., Mori Y., Sasaki T., Sasaki A., Nagatsuma T., Ochiai Y., Fukasaku N., *Jpn. J. Appl. Phys.*, **41**, L1028 (2002).
4) Ogawa T., Okamoto M., Yagi H., Mori Y., Hatta A., Ito T., Sasaki T., Hiraki A., *Diamond Films Technol.*, **6**, 87–94 (1996).
5) Ogawa T., Okamoto M., Khin Y. Y., Mori Y., Hatta A., Ito T., Sasaki T., Hiraki A., *Diamond Related Mater.*, **6**, 1015–1018 (1997).
6) Okamoto M., Mori Y., Sasaki T., *Jpn. J. Appl. Phys.*, **38**, 2114–2115 (1999).
7) Yap Y. K., Kida S., Aoyama T., Mori Y., Sasaki T., *Jpn. J. Appl. Phys.*, **37**, L746–L748 (1998).
8) Garetz B. A., Aber J. E., Goddard N. L., Young R. G., Mayerson A. S., *Phys. Rev. Lett.*, **77**, 3475 (1996).
9) Mori Y., Kuroda I., Nakajima S., Sasaki T., Nakai S., *Appl. Phys. Lett.*, **67**, 1818 (1995).
10) Sasaki T., Mori Y., Kuroda I., Nakajima S., Yamaguchi K., Watanabe S., *Acta Crystal-*

logr, **C51**, 2222 (1995).
11) Mori Y., Sasaki T., *Ouyoubutsuri*, **66**, 965 (1997).
12) Nishioka M., Kawamura F., Yoshimura M., Mori Y., Sasaki T., Proceedings of the Conference on Lasers and Electro-Optics, CTuF2, 2003.
13) Adachi H., Takano K., Hosokawa Y., Inoue T., Mori Y., Matsumura H., Yoshimura M., Tsunaka Y., Morikawa M., Kanaya S., Masuhara H., Kai Y., Sasaki T., *Jpn. J. Appl. Phys.*, **42**, L798–L800 (2003).
14) Adachi H., Niino A., Kinoshita T., Warizaya M., Maruki R., Takano K., Matsumura H., Inoue T., Murakami S., Mori Y., Sasaki T., *J. Biosci. Bioeng.*, **101**, 83 (2006).
15) Murakami S., Nakashima R., Yamashita E., Yamaguchi A., *Nature*, **419**, 587 (2002).
16) Adachi H., Murakami S., Niino A., Matsumura H., Takano K., Inoue T., Mori Y., Yamaguchi A., Sasaki T., *Jpn. J. Appl. Phys.*, **43**, 10B, L1376–L1378 (2004).
17) Tsukazaki T., Mori H., Fukai S., Numata T., Perederina A., Adachi H., Matsumura H., Takano K., Murakami S., Inoue T., Mori Y., Sasaki T., Vassylyev D. G., Nureki O., Ito K., *Acta Crystallogr.*, **F62**, 376 (2006).
18) Numata T., Ikeuchi Y., Fukai S., Adachi H., Matsumura H., Takano K., Murakami S., Inoue T., Mori Y., Sasaki T., Suzuki T., Nureki O., *Acta Crystallogr.*, **F62**, 368 (2006).
19) Numata T., Ikeuchi Y., Fukai S., Suzuki T., Nureki O., *Nature*, **442**, 419 (2006).
20) Fukunishi Y., Mikami Y., Kubota S., Nakamura H., *J. Mol. Graph. Model.*, **25**, 61–70 (2005).
21) Fukunishi Y., Mikami Y., Takedomi K., Yamanouchi M., Shima H., Nakamura H., *J. Med. Chem.*, **49**, 523?533 (2006).
22) Krungkrai S. R., Tokuoka K., Kusakari Y., Inoue T., Adachi H., Matsumura H., Takano K., Murakami S., Mori Y., Kai Y., Krungkrai J., Horii T., *Acta Crystallogr.*, **F62**, 542–545 (2006).
23) Tokuoka K., Kusakari Y., Krungkrai S. R., Matsumura H., Kai Y., Krungkrai J., Horii T., Inoue T., J. Biochem., **143**(1), 67 (2008).

Published online 5 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* **D907–D912**
doi:10.1093/nar/gkm948

GLIDA: GPCR—ligand database for chemical genomics drug discovery—database and tools update

Yasushi Okuno[1,*], Akiko Tamon[2], Hiroaki Yabuuchi[1], Satoshi Niijima[1], Yohsuke Minowa[1], Koichiro Tonomura[1], Ryo Kunimoto[1] and Chunlai Feng[1]

[1]Department of PharmacoInformatics, Center for Integrative Education of Pharmacy Frontier, Graduate School of Pharmaceutical Sciences, Kyoto University and [2]Bio Science Group, IT Solution Div.1, Industry Solution Business Unit, Mitsui Knowledge Industry, Osaka city, Japan

Received September 6, 2007; Revised October 14, 2007; Accepted October 15, 2007

ABSTRACT

G-protein coupled receptors (GPCRs) represent one of the most important families of drug targets in pharmaceutical development. GLIDA is a public GPCR-related Chemical Genomics database that is primarily focused on the integration of information between GPCRs and their ligands. It provides interaction data between GPCRs and their ligands, along with chemical information on the ligands, as well as biological information regarding GPCRs. These data are connected with each other in a relational database, allowing users in the field of Chemical Genomics research to easily retrieve such information from either biological or chemical starting points. GLIDA includes a variety of similarity search functions for the GPCRs and for their ligands. Thus, GLIDA can provide correlation maps linking the searched homologous GPCRs (or ligands) with their ligands (or GPCRs). By analyzing the correlation patterns between GPCRs and ligands, we can gain more detailed knowledge about their conserved molecular recognition patterns and improve drug design efforts by focusing on inferred candidates for GPCR-specific drugs. This article provides a summary of the GLIDA database and user facilities, and describes recent improvements to database design, data contents, ligand classification programs, similarity search options and graphical interfaces. GLIDA is publicly available at http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/. We hope that it will prove very useful for Chemical Genomics research and GPCR-related drug discovery.

INTRODUCTION

The family of G-protein coupled receptors (GPCRs) represents one of the most important classes of pharmaceutical targets (1). Among the more than 1000 GPCRs encoded in the human genome, more than 400 are of potential therapeutic interest (2). Currently the drugs available on the market address only 30 GPCRs, which represent a small fraction of the GPCR target family. A large majority of human-derived GPCRs still remain promising drug targets, and thus a key goal of GPCR research related to drug design is to identify new ligands for such target GPCRs.

With the unprecedented accumulation of genomic information, databases and bioinformatics have become essential tools to guide GPCR research (3). The GPCRDB (2) and IUPHAR receptor database (IUPHAR-RD) (4) are representatives of widely used public databases covering GPCRs. These databases, which provide substantial data on the GPCR proteins and pharmacological information on receptor proteins containing GPCRs, are mainly focused on biological aspects of the GPCR gene products or proteins. In spite of the significance of ligand compounds as drug leads, the relationships between GPCRs and their ligands and/or chemical information on the ligands themselves are not yet fully covered.

On the other hand, there is increasing interest in publicly collecting and applying chemical as well as biological information in the post-genome era (5–8). This new trend is called 'Chemical Genomics', and it aims to identify all possible chemical ligands and drugs for all targets families (9,10). There is a vast amount of information on the interactions between small molecules and proteins/genes. However, compound–protein interactions have not yet been analyzed on a large scale, and there are no effective methods to extract meaningful

*To whom correspondence should be addressed. Tel: +81 75 753 4559; Fax: +81 75 753 4544; Email: okuno@pharm.kyoto-u.ac.jp

information from the data in a comprehensive manner. Therefore, we need to integrate chemoinformatics and bioinformatics into a common computational platform for mining of Chemical Genomics data (11).

GLIDA (GPCR-Ligand DAtabase) is a public GPCR-related Chemical Genomics database designed to simultaneously mine biological information on GPCRs and chemical information on their ligands. It provides various analytical data regarding GPCR–ligand correlations by incorporating bioinformatics and chemoinformatics techniques, and thus it should prove very useful for GPCR-related drug discovery from the viewpoint of Chemical Genomics research. There have been several major improvements to GLIDA since it was last described in Ref. (12): (i) there are more increments in the entries of the ligands and the corresponding ligand–GPCR pairs; (ii) the ligands are originally classified using a new strategy; (iii) additional options are available within the similarity search program for the GPCRs and ligands and (iv) the graphical interface to display the correlation maps between GPCRs and ligands has been enhanced.

DATA CONTENTS

GLIDA contains three types of primary data: biological information on GPCRs, chemical information on their ligands and information on binding of the GPCR–ligand pairs. The GPCR entries were acquired from human, mouse and rat entries deposited in the GPCRDB because these three species include sufficient information regarding ligands, and rats and mice are representative model animals used in drug discovery research. The ligand-binding information was manually collected and curated using various public web sites and commercial databases such as the IUPHAR-RD, PubMed (5), PubChem (5), DrugBank (13), Ki Database (14) and MDL ISIS/Base 2.5. Table 1 indicates the size and scope of the GLIDA database. In particular, we have dramatically expanded the entry number of ligands and the corresponding ligand–GPCR pairs. The latest GLIDA version includes 24 077 ligand entries and 39 140 GPCR–ligand pair entries, representing nearly 35-fold and 20-fold increases, respectively, since the last publication of GLIDA in 2006. The total number of GPCR entries remains unchanged, but entries with associated ligand information have increased slightly, suggesting that it is difficult to de-orphan the GPCRs whose ligands have not yet been identified (15).

GPCR and ligand data

The database lists general information on GPCR and ligand data, respectively. The general information table listing GPCRs contains gene names, family names, protein sequences (in fasta format) and links to other biological databases, such as GPCRDB, UniProt (16), IUPHAR-RD, Entrez Gene (17) and KEGG (18). The ligand result page provides a general information table containing names, molecular structures, CAS registry numbers, formulas, molecular weights, structure files and links to PubChem, KEGG, ChEBI (8) and DrugBank that are in publicly available chemical databases.

Table 1. The current numbers of GLIDA ligands and GPCRs and their respective links.

Information item	Number of entries
GPCR entries	3738
Links to Entrez Gene	3073
Links to GPCRDB	3738
Links to UniProt	3738
Links to IUPHAR	446
Links to KEGG	595
Ligand entries	24 077
Cas registry number	2425
Molecular structure	23 216[a]
Links to PubChem	1821
Links to ChEBI	103
Links to KEGG	664
Links to DrugBank	479
Cluster number	300[b]
GPCR–ligand pair entries	39 140
GPCR entries	410
Ligand entries	24 077
Activity	
Agonist	8305
Full Agonist	2325
Partial Agonist	262
Antagonist	28 132
Inverse Agonist	116

[a]Molecular structures consist of MDL MOL files and original files converted into KEGG atom types. The numbers of MDL MOL files and KEGG-type files are 23 216 and 23 214, respectively. PCA calculation was performed for 23 214 KEGG-type files.
[b]This cluster number (300) is different from the number of the selected principal components (314). No compounds were assigned to 14 principal components.

Information on binding of GPCR–ligand pairs

The interaction information relating GPCRs to particular ligands, a key issue for GPCR-related drug discovery, is deposited in a relational database. GLIDA allows users to retrieve GPCR–ligand-binding information dynamically and continuously. When users retrieve a GPCR (or ligand) entry, its result page displays all entries showing the corresponding ligands (or GPCR) entries with their binding activity types, as well as references. The references are hyperlinked with the corresponding PubMed literature. The activity types include agonist, antagonist and full, partial or inverse agonist (Table 1). Here the detail annotations such as full, partial or inverse agonist are not finished yet. The ligands classified as agonists are possible full agonists or partial agonists. Inverse agonists can be also contained among the antagonists.

WEB INTERFACE AND APPLICATION

GLIDA is available at http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/. The web interface of GLIDA includes a GPCR search page (Figure 1a) and a ligand search page (Figure 1b). Each page consists of a classification menu and a keyword search box. The users can search a GPCR (or ligand) manually using the classification tool,

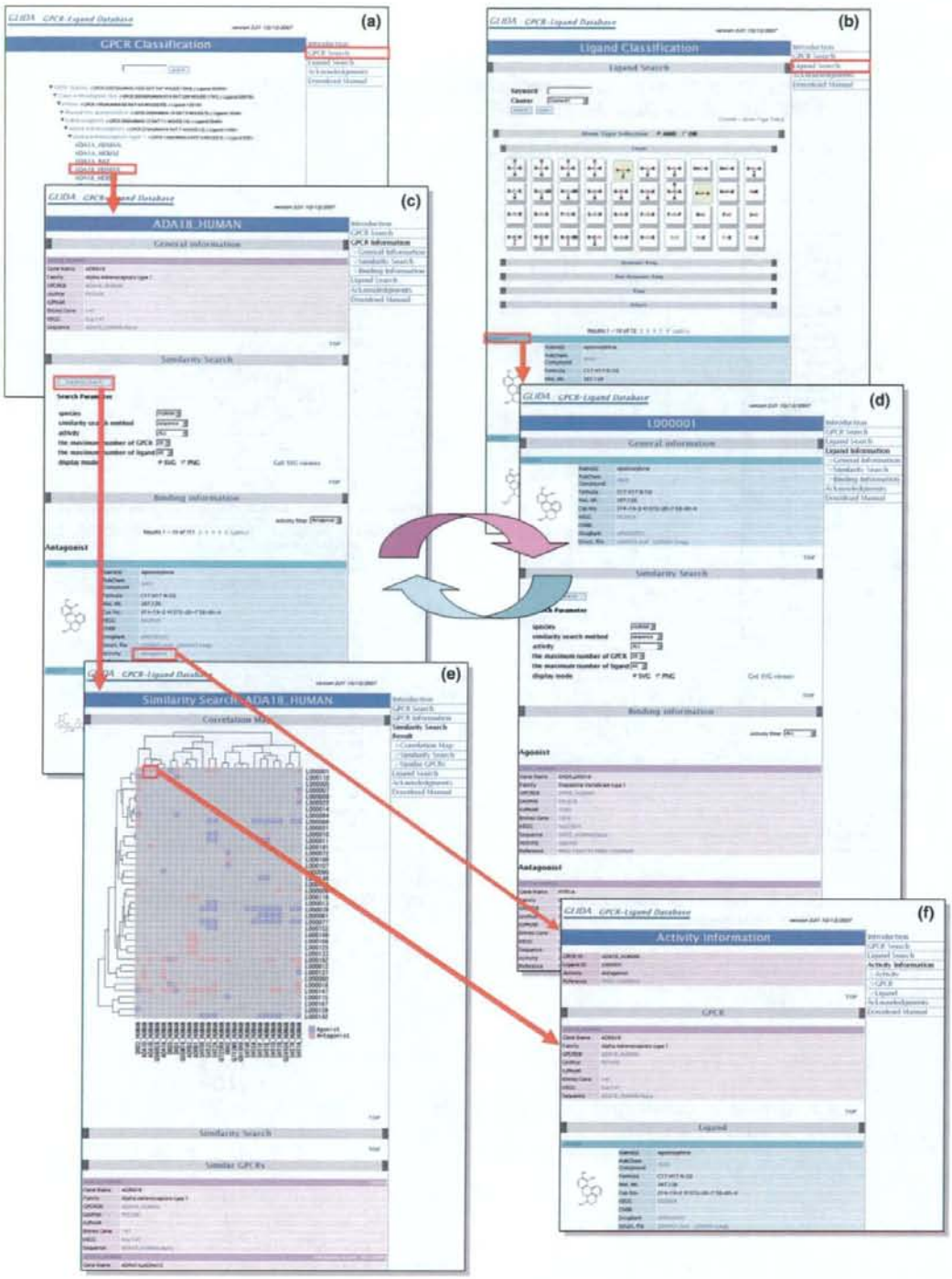

Figure 1. A screenshot of GLIDA showing linked relations among search pages (**a** and **b**), result pages (**c** and **d**), an analytical report page (**e**), and a binding information page (**f**). The analytical report page consists of a correlation map and a list resulting from a similarity search. Red and blue colors of the spots on the correlation map indicate the ligand activities of antagonists including inverse agonist and agonists including full/partial agonist, respectively.

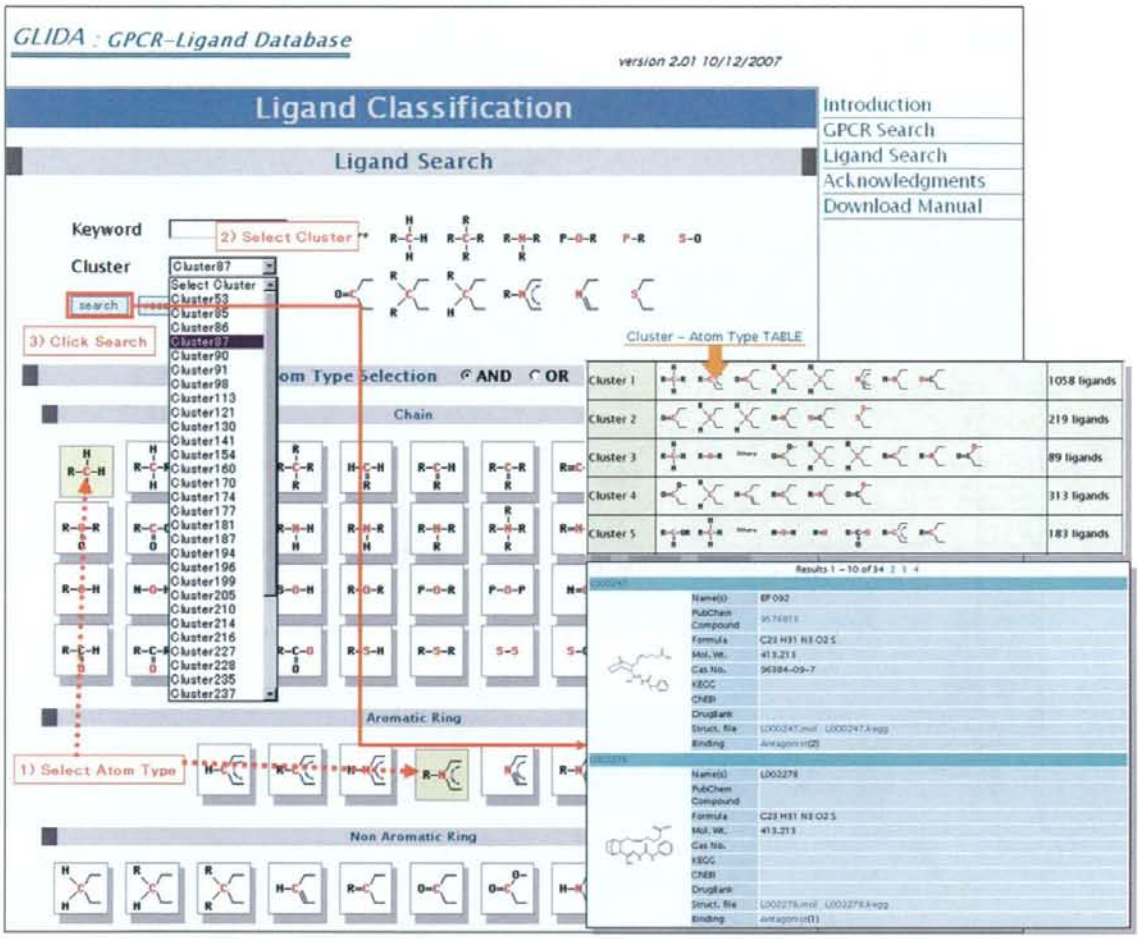

Figure 2. A screenshot of the ligand search process on the ligand classification page. Users can search the ligands from two starting points: keyword search and cluster selection. If they have a chemical structure of their query compound, the ligand search is performed using the cluster selection tool as follows. Selecting a set of atom types (step 1) that the query compound contains, the pull-down menu of cluster selection displays the list of the only clusters that include selected atom types as the principal components (step 2). By selecting a cluster from the list, users can check the principal component's atoms on the upper right section of the page. Finally, upon clicking the search button, GLIDA displays the list of all ligands classified in the selected cluster (step 3). The 'Atom Type TABLE' button links the user to the page showing the cluster size and representative atom types for each cluster.

or automatically by using the keyword search function. Every GPCR (or ligand) has its own results page (Figure 1c or d) containing a general information table regarding a GPCR (or ligand), a table of its correlated ligands (or GPCRs) and a menu button to carry out a similarity search and correlation analysis.

Classification of GPCRs and ligands

The GPCR classification table on the search page was adapted from the phylogenetic tree within the GPCRDB information system (http://www.gpcr.org/7tm/phylo/phylo.html). The GPCR classification table displays the entries of the corresponding GPCRs at the tree branches, and these are hyperlinked with the corresponding result pages (Figure 1a). GLIDA also provides an original ligand classification (Figures 1b and 2). With the great increase in ligand entries, we have to improve our method of classifying all the ligands in GLIDA. Hierarchical clustering and its tree representation, which were used in the old version of GLIDA, are unsuitable for the data mining of huge chemical databases. We therefore have adopted principal component analysis (PCA) for clustering of 23 214 ligand structures in this new version, as follows. We generated frequency profiles of the atoms and the bonds converted into the KEGG atom types from MDL MOL files of ligand entries (19). The KEGG-type profile for each ligand is shown in 'Struct. file' item of

general information table of GLIDA. PCA was applied to the data matrix consisting of 700 KEGG-type features' columns and 23 214 ligand entries' rows. The resulting principal components (PCs) constitute a new set of linearly independent, orthogonal axes that capture the directions of maximum variance in the data. The samples (chemical compounds) were then projected onto these PC axes. Herein, we used the top 314 PCs as seeds of clusters that account for >80% (cumulative proportion) of the total variance. Finally, each compound was assigned to the PC cluster having the maximum score among the 314 PCs. In order to annotate the features of each cluster (PC), we selected for each PC the atom types and their bonds corresponding to the top 10 loadings having the largest magnitude. The ligand classification page displays a table of all the atom types selected by PCA (Figure 2). By clicking on some of the atoms in this table, users can search clusters that include the selected atom types. Consequently, the ligands relevant to users' interests are included in the retrieved cluster.

Similarity search and correlation map for GPCRs and ligands

The fact that similar proteins bind similar ligands is the underlying principle of the Chemical Genomics approach to drug discovery (11). GLIDA has a variety of similarity search functions for GPCRs and ligands, respectively, on its result pages (Figure 1c or d). Alignment scores for protein sequences generated by the BLAST algorithm provide similarity measures for GPCRs. In addition to sequence similarity, gene expression patterns in tissue origins and developmental stages were used as similarity measures. The expression data for each GPCR was generated from the EST sequences in different libraries served from NCBI/Unigene (http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi). We can thereby retrieve the GPCRs that present tissue-/stage-specific distribution similar to a query GPCR. For example, co-expression information on specific GPCRs enables us to speculate about GPCR-heterodimerization that might have an effect on their activity (1). Ligand similarity is defined by the dissimilarity (distance) of frequency profile patterns generated from the constitutive atoms and bonds of the chemical structure, using the KEGG atom types (19,20). From the similarity search, the most similar GPCRs (or ligands) within the users' selected parameters are retrieved and listed with their similarity scores on an analytical report page (Figure 1e). In the latest GLIDA version, various parameters have been added as search options, such as selections of species, ligand activities, displayed number of GPCRs/ligands and map graphical mode. As another result of similarity search calculations, GLIDA illustrates the correlation map (Figure 1e) showing homologous GPCRs (or ligands) and their ligands (or GPCRs) that are retrieved. This map shows spots that match the GPCRs and their ligands in a 2D matrix. The ordering along the x-axis and the y-axis are calculated respectively by two-way clustering of the GPCRs and the ligands, based on their similarities. In particular, the ordering along the x- and y-axes allows users to evaluate the sequence similarities among GPCRs and the correlation coefficients among ligands simultaneously. By analyzing the correlation patterns between GPCRs and ligands that are illustrated by these maps, we can gain detailed knowledge about their interactions. We can then utilize this information to infer possible candidates for development of GPCR-specific drugs. Furthermore, we have enhanced a graphical interface to display the correlation map between GPCRs and ligands. Graphics are an important tool to aid visualization and interpretation of high-dimensional data. The old version of GLIDA used only the PNG (Portable Network Graphics) format to display a GPCR–ligand correlation map. Due to the great increase in entries, the latest GLIDA version introduces the SVG (Scalable Vector Graphics) format, which is adaptable to an enormous correlation map size. The SVG vector image can be scaled indefinitely without loss of image quality, while the PNG bitmap image cannot. Users must install the free plug-in software on their computer in advance to use the SVG format (http://www.adobe.com/svg/viewer/install/). In the case of uninstalled devices, PNG representation should be selected as a graphical mode. Figure 1 shows an example of the GPCR–ligand search and analysis process starting from a GPCR query using GLIDA.

DISCUSSION AND FUTURE DIRECTIONS

GLIDA provides a unique database useful for GPCR-related Chemical Genomics research and drug discovery. GLIDA is distinct from other public Chemical Genomics databases because it contains original, GPCR-specific chemical entries and offers a common mining platform of bioinformatics and chemoinformatics. GLIDA provides several advantages over other databases, in that a search can be started either from a GPCR or from a ligand. Thus, searches can be carried out in a dynamic and user-friendly way. GLIDA's coverage of chemical and biological information simultaneously also provides an advantage to users by saving them the time and labor required to search multiple databases. The ligand search page is another distinct characteristic of GLIDA, in that it displays the structural distribution of ligands. It thereby facilitates research on GPCR-related drugs by incorporating structural aspects of the ligand compounds into the search. The analytical report pages resulting from the calculated structural similarities of GPCRs and ligands can give the user deep insights into the GPCR–ligand relationships. The lists of neighboring ligands (or GPCRs) and the correlation maps are useful visualization tools for analyzing correlations among the structural features and the GPCR–ligand-binding properties. Because this database system can be applied to proteins other than the GPCR family, it may also be considered as a promising database for other types of Chemical Genomics research. One critical issue is how to define similarity metrics for proteins and ligands, because the underlying principle of GLIDA is that similar receptors bind similar ligands. For example, ligand similarity can be defined by manifold representations such as graph, fingerprint and descriptors.

Protein similarity can be also measured in different ways such as overall sequence homology (phylogenetic relationships), consensus motifs, common binding sites, 3D structures and reported functional annotations. Therefore we will add new menus incorporating these various similarity metrics for GPCRs and ligands. GLIDA will be updated continuously. In particular, we are now planning to add the drawing tool of chemical structures and to expand the ligand-searching function for an arbitrary chemical query.

ACKNOWLEDGEMENTS

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, from the JSPS, KAKENHI, Grant-in-Aid for Publication of Scientific Research Results and from the Ministry of Health, Labour and Welfare of Japan. Financial support from the SUNTORY INSTITUTE FOR BIOORGANIC RESEARCH, the TATEISI SCIENCE AND TECHNOLOGY FOUNDATION and the Okawa Foundation for Information and Telecommunications is gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. George,S.R., O'Dowd,B.F. and Lee,S.P. (2002) G-protein-coupled receptor oligomerization and its potential for drug discovery. *Nature Rev. Drug Discov.*, **1**, 808–820.
2. Horn,F., Bettler,E., Oliveira,L., Campagne,F., Cohen,F.E. and Vriend,G. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.
3. Strachan,R., Ferrara,G. and Roth,BL. (2006) Screening the receptorome: an efficient approach for drug discovery and target validation. *Drug Discov. Today*, **11**, 708–716.
4. Foord,S.M., Bonner,T.I., Neubig,R.R., Rosser,E.M., Pin,J.P., Davenport,A.P., Spedding,M. and Harmar,A.J. (2005) International Union of Pharmacology. XLVI. G Protein-Coupled Receptor List. *Pharmacol. Rev.*, **57**, 279–288.
5. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, 12.
6. Schreiber,SL. (2004) Stuart Schreiber: biology from a chemist's perspective. Interview by Joanna Owens. *Drug Discov. Today*, **9**, 299–303.
7. Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, D402–D404.
8. Brooksbank,C., Cameron,G. and Thornton,J. (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.*, **33**, D46–D53.
9. Zerhouni,E. (2003) The NIH Roadmap. *Science*, **302**, 63–72.
10. Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.
11. Klabunde,T. (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.*, **152**, 5–7.
12. Okuno,Y., Yang,J., Taneishi,K., Yabuuchi,H. and Tsujimoto,G. (2006) GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.*, **34**, D673–D677.
13. Wishart,D.S., Knox,C., Guo,A.C., Shrivastava,S., Hassanali,M., Stothard,P., Chang,Z. and Woolsey,J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
14. Roth,B.L., Lopez,E., Beischel,S., Westkaemper,R.B. and Evans,J.M. (2004) Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol. Ther.*, **102**, 99–110.
15. Civelli,O. (2005) GPCR deorphanizations: the novel, the known and the unexpected transmitters. *Trends Pharmacol. Sci.*, **26**, 15–19.
16. The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Research*, **35**, D193–D197.
17. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
18. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
19. Hattori,M., Okuno,Y., Goto,S. and Kanehisa,M. (2003) Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.
20. Kotera,M., Okuno,Y., Hattori,M., Goto,S. and Kanehisa,M. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.

Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection

Satoshi Niijima and Yasushi Okuno

Abstract— **Until recently, numerous feature selection techniques have been proposed and found wide applications in genomics and proteomics. For instance, feature/gene selection has proven to be a powerful tool for biomarker discovery from microarray and mass spectrometry data. While supervised feature selection has been explored extensively, there are only a few unsupervised methods that can be applied to exploratory data analysis in which class information is unavailable. In this paper, we address the problem of unsupervised feature selection. First, we extend Laplacian linear discriminant analysis (LLDA) to unsupervised cases. Second, we propose an efficient algorithm for computing LLDA. Finally, a new unsupervised feature selection algorithm, called LLDA-based Recursive Feature Elimination (LLDA-RFE), is presented. We apply LLDA-RFE to several public datasets of cancer microarrays and compare its performance with those of state-of-the-art unsupervised methods, Laplacian score and SVD-entropy, and of a supervised filter method, Fisher score. Our results demonstrate that LLDA-RFE outperforms Laplacian score and shows favorable performance against SVD-entropy. It performs even better than Fisher score for some of the datasets, despite the fact that LLDA-RFE is fully unsupervised.**

Index Terms— **Feature selection, linear discriminant analysis, graph Laplacian, microarray data analysis.**

I. Introduction

In recent years, feature/gene selection methods have been widely used in genomics and proteomics to handle a deluge of data produced by high-throughput technologies such as microarray and mass spectrometry. In microarray studies, for instance, a small fraction of genes typically exhibit significant differential expression among tens of thousands of genes whose expression levels are measured simultaneously. Thus, it is of great importance to identify genes relevant to a biological phenomenon of interest and to characterize their expression profiles. Gene selection can be useful for multiple purposes: to save computational costs of subsequent analysis by reducing the number of genes; to improve the prediction performance of classifiers by using discriminative genes only; and to identify informative genes for further investigation of their biological relevance. Specifically, gene selection has proven to be a powerful tool for biomarker discovery, i.e. searching for potential marker genes contributing to classification of cancer subtypes or prediction of clinical outcomes, which leads to more reliable diagnosis and better treatments of cancer.

To date, numerous techniques for feature selection have been developed [12] and also been applied successfully to the analysis of biological data with many features. In contrast to supervised feature selection, however, unsupervised feature selection has not yet been explored extensively. Indeed, there have been only a few unsupervised methods proposed until recently [7], [14], [28], [30]. Unsupervised feature selection is of great use particularly for class discovery, where class information is unavailable. For instance, clustering is usually performed to find clusters in microarray samples on the basis of the expression profiles of all genes, but the clusters so obtained can be obscured by the large number of irrelevant genes. Therefore, unsupervised feature selection is essential to the exploratory analysis of biological data. Moreover, even when class labels are provided by external knowledge, but may be unreliable or mislabeled, overfitting can be alleviated by performing feature selection in an unsupervised manner. It is obviously more challenging to identify features that reveal underlying cluster structures in the samples than to find those exhibiting similar patterns across all the samples.

To address this problem, we propose a new unsupervised feature selection method, called Laplacian linear discriminant analysis-based Recursive Feature Elimination (LLDA-RFE). LLDA-RFE is closely related to Laplacian score [15], which is also based on graph Laplacian and can be applied in an unsupervised manner. The major difference is that, whereas Laplacian score is a univariate approach, LLDA-RFE is multivariate, allowing for selecting features that contribute to discrimination in combination with other features. Recently, Wolf and Shashua [30] proposed the $Q-\alpha$ algorithm, taking advantage of the spectral properties of the graph Laplacian of features. While the $Q-\alpha$ algorithm has an interesting property that the sparsity of features naturally emerges, it does not scale well to the feature size. Also, it involves iterative computations on a matrix of the feature size in a least-squares optimization process to ensure a local maximum solution. In contrast, our proposed algorithm for LLDA-RFE is computationally tractable and has a global maximum solution. It is shown that LLDA includes the maximum margin criterion (MMC) [18] as a supervised case. Although LLDA-RFE is a natural extension of MMC-RFE, the proposed algorithm need not reduce dimensionality before applying LLDA, unlike the MMC-RFE algorithm proposed previously [23].

We compare the performance of LLDA-RFE with those of state-of-the-art unsupervised feature selection methods, Laplacian score and SVD-entropy [28], on seven public datasets of cancer microarrays. The performances of these methods are evaluated by their capability of identifying discriminative genes without using class information. We also compare the performance between LLDA-RFE and a supervised filter method, Fisher score [8], [15]. Experimental results demonstrate that LLDA-RFE outperforms Laplacian score and shows favorable performance against SVD-entropy. Despite the fact that LLDA-RFE is fully unsupervised, it performs even better than Fisher score for some of the datasets.

The rest of this paper is organized as follows: In Section II, we give outlines of LDA and the MMC. We then introduce LLDA and extend it to unsupervised cases in Section III. An efficient algorithm for LLDA is also proposed. We present the LLDA-RFE

Manuscript received XXX XX, 2007; revised XXX XX, 2007.

S. Niijima and Y. Okuno are with the Department of PharmacoInformatics, Frontier Education Center, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan.

algorithm for feature selection in Section IV. Section V describes related work on unsupervised feature selection. Experimental results on seven microarray datasets are presented and discussed in Section VI. Finally, we give concluding remarks in Section VII.

II. LDA AND MMC

Linear discriminant analysis (LDA) aims to find a set of projection vectors that maximize the between-class scatter and simultaneously minimize the within-class scatter, thereby achieving maximum discrimination [9].

Let $\boldsymbol{X} \in \mathbb{R}^{p \times n}$ be a sample matrix containing $\boldsymbol{x}_i, i = 1, \ldots, n$ as columns, where n is the number of samples and p is the number of features. The between-class scatter matrix $\boldsymbol{S}_b$ and the within-class scatter matrix $\boldsymbol{S}_w$ are defined as:

$$\boldsymbol{S}_b = \frac{1}{n}\sum_{k=1}^{c} n_k (\boldsymbol{m}^{(k)} - \boldsymbol{m})(\boldsymbol{m}^{(k)} - \boldsymbol{m})^T,$$

$$\boldsymbol{S}_w = \frac{1}{n}\sum_{k=1}^{c}\sum_{j=1}^{n_k} (\boldsymbol{x}_j^{(k)} - \boldsymbol{m}^{(k)})(\boldsymbol{x}_j^{(k)} - \boldsymbol{m}^{(k)})^T,$$

where c is the number of classes; n_k is the number of samples in class k; $\boldsymbol{x}_j^{(k)}$ is the jth sample in class k; $\boldsymbol{m}^{(k)}$ and $\boldsymbol{m}$ are the mean vector of class k and the total mean vector, respectively. Then, classical LDA finds the projection matrix $\boldsymbol{W}$ by maximizing the Fisher criterion

$$J_{LDA}(\boldsymbol{W}) = trace\big((\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W})^{-1}(\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W})\big). \quad (1)$$

By solving a generalized eigenvalue problem, $\boldsymbol{W}$ can be found as the eigenvectors of $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ corresponding to the largest eigenvalues. However, when the dimensionality of samples is larger than the sample size, i.e. $p > n$, $\boldsymbol{S}_w$ becomes singular and we cannot compute $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$, which is a major drawback of classical LDA. This is known as the singularity problem or the small sample size problem.

To overcome this problem, Li *et al.* [18] recently proposed to use the maximum margin criterion (MMC) instead of (1) to find the projection vectors. The MMC is defined as:

$$J_{MMC}(\boldsymbol{W}) = trace\big(\boldsymbol{W}^T(\boldsymbol{S}_b - \boldsymbol{S}_w)\boldsymbol{W}\big). \quad (2)$$

In this case, the projection matrix $\boldsymbol{W}$ that maximize the criterion (2) can be found as the eigenvectors of $\boldsymbol{S}_b - \boldsymbol{S}_w$ corresponding to the largest eigenvalues. Li *et al.* proposed an efficient algorithm to compute the projection matrix of the MMC under the constraint that $\boldsymbol{W}^T\boldsymbol{S}_t\boldsymbol{W} = \boldsymbol{I}$, where $\boldsymbol{S}_t$ is the total scatter matrix. This is found to be the same as the uncorrelated LDA (ULDA) algorithm in [32]. Also, an efficient algorithm for the MMC subject to the orthogonality constraint on $\boldsymbol{W}$, i.e. $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$, was presented in [23]. In both cases, we need not compute the inverse of $\boldsymbol{S}_w$, hence the singularity problem can be easily avoided.

It should be noted that the MMC is not equivalent to the Fisher criterion. As Loog [21] disproved the theorem in [26], the discriminant vectors obtained by maximizing (2) are not generally the same as those obtained by maximizing (1). More precisely, although ULDA can be considered as an extension of classical LDA to small sample size cases [32], the MMC with the orthogonality constraint does not necessarily yield projection vectors that are optimal for discrimination. In practice, a better discrimination can be achieved by balancing the between-class and within-class scatters using the following criterion as in [20]:

$$J_{MMC}(\boldsymbol{W}) = trace\big(\boldsymbol{W}^T(\boldsymbol{S}_b - \mu\boldsymbol{S}_w)\boldsymbol{W}\big), \quad (3)$$

where μ is a non-negative constant. It is clear that (2) is a special case of (3). In the following sections, we focus on the MMC defined by (2) with the orthogonality constraint.

III. UNSUPERVISED LLDA

A. Extension of LLDA to unsupervised cases

We can write the total and within-class scatter matrices as follows:

$$\begin{aligned}\boldsymbol{S}_t &= \frac{1}{n}\boldsymbol{X}\big(\boldsymbol{I} - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^T\big)\boldsymbol{X}^T \\ &= \frac{1}{n}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{W}_g)\boldsymbol{X}^T, \\ \boldsymbol{S}_w &= \frac{1}{n}\boldsymbol{X}(\boldsymbol{I} - \sum_{k=1}^{c}\frac{1}{n_k}\boldsymbol{e}^{(k)}\boldsymbol{e}^{(k)^T})\boldsymbol{X}^T \\ &= \frac{1}{n}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{W}_\ell)\boldsymbol{X}^T,\end{aligned}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{e} = (1, 1, \ldots, 1)^T$ is an n-dimensional vector, and $\boldsymbol{e}^{(k)}$ is an n-dimensional vector with $e_i^{(k)} = 1$ if $\boldsymbol{x}_i$ belongs to class k, and 0 otherwise. In terms of graph Laplacians [6], $\boldsymbol{I} - \boldsymbol{W}_g$ can be viewed as the global Laplacian of a graph such that all vertices are connected with a constant weight of $1/n$, and $\boldsymbol{I} - \boldsymbol{W}_\ell$ as the local Laplacian of a graph such that vertices are connected with a constant weight of $1/n_k$ only when both belong to the kth class.

From the relationship [9]

$$\boldsymbol{S}_t = \boldsymbol{S}_b + \boldsymbol{S}_w,$$

it follows that

$$\begin{aligned}\boldsymbol{S}_b - \boldsymbol{S}_w &= \boldsymbol{S}_t - 2\boldsymbol{S}_w \\ &= \frac{1}{n}\boldsymbol{X}\big((\boldsymbol{I} - \boldsymbol{W}_g) - 2(\boldsymbol{I} - \boldsymbol{W}_\ell)\big)\boldsymbol{X}^T. \quad (4)\end{aligned}$$

The MMC represented in this form is referred to as Laplacian linear discriminant analysis (LLDA) in [26] and was applied to extract discriminant features in supervised scenarios.

In this study, we extend (4) to unsupervised cases. We first define the global similarity matrix $\boldsymbol{K}_g$ and the local similarity matrix $\boldsymbol{K}_\ell$ as:

$$[\boldsymbol{K}_g]_{ij} = \begin{cases} k(\boldsymbol{x}_i, \boldsymbol{x}_j), & \text{if } i \neq j \\ 0, & \text{otherwise,} \end{cases}$$

$$[\boldsymbol{K}_\ell]_{ij} = \begin{cases} k(\boldsymbol{x}_i, \boldsymbol{x}_j), & \text{if } \boldsymbol{x}_i \text{ is among } k_\ell \text{ nearest neighbors of } \boldsymbol{x}_j \\ & \text{or } \boldsymbol{x}_j \text{ is among } k_\ell \text{ nearest neighbors of } \boldsymbol{x}_i \\ 0, & \text{otherwise.} \end{cases}$$

Here, $k(\cdot, \cdot)$ represents the similarity between each pair of samples, and the standard measures include heat kernel (Gaussian kernel), inner product and Euclidean distance. In supervised cases, prior class information can also be reflected to guide the graph construction [31]. Let $\boldsymbol{L}_g$ and $\boldsymbol{L}_\ell$ be the normalized global and local Laplacian matrices, respectively as:

$$\boldsymbol{L}_g = \boldsymbol{I} - \boldsymbol{D}_g^{-\frac{1}{2}}\boldsymbol{K}_g\boldsymbol{D}_g^{-\frac{1}{2}},$$

$$\boldsymbol{L}_\ell = \boldsymbol{I} - \boldsymbol{D}_\ell^{-\frac{1}{2}}\boldsymbol{K}_\ell\boldsymbol{D}_\ell^{-\frac{1}{2}},$$

where D_g and D_ℓ are diagonal matrices such that $[D_g]_{ii} = \sum_j [K_g]_{ji}$ and $[D_\ell]_{ii} = \sum_j [K_\ell]_{ji}$. Then, we seek to find a set of projection vectors W that maximize the following criterion:

$$J_{LLDA}(W) = trace\big(W^T(S_g - 2S_\ell)W\big), \qquad (5)$$

where S_g and S_ℓ are the global and local scatter matrices defined as:

$$S_g = \frac{1}{n} X L_g X^T,$$
$$S_\ell = \frac{1}{n} X L_\ell X^T.$$

It is easy to check that, when we set $[K_g]_{ij} = 1/n$ for all i, j; and $[K_\ell]_{ij} = 1/n_k$ if x_i and x_j are both in the kth class, and 0 otherwise, L_g and L_ℓ respectively become $I - W_g$ and $I - W_\ell$, hence (5) includes the MMC as a special case.

In general, (5) does not require class information and can be used in an unsupervised manner. The construction of the local scatter matrix is based on the assumption that, if x_i and x_j are close, they are likely to belong to the same cluster. Under the condition that class labels are unavailable, we cannot explicitly consider the separability of different clusters, which is represented by the between-class scatter in classical LDA and the MMC. In the objective function (5), it is implicitly represented by the difference between the global scatter and the local scatter. Therefore, discriminative features can be extracted even in unsupervised scenarios. In this paper, we refer to unsupervised LLDA simply as LLDA.

Note that the reason for using the normalized graph Laplacians is that the criterion (5) without normalization may be affected by the scale of the similarity measure or by the choice of the number of nearest neighbors, since (5) is defined as the difference rather than the ratio of the global scatter to the local scatter. Also, the use of normalized graph Laplacian is known to be effective in spectral clustering (e.g. [24]).

B. *Efficient algorithm for LLDA*

Similarly to the case of (2), the projection matrix W that maximize the criterion (5) subject to the orthogonality constraint can be found as the eigenvectors of $S_g - 2S_\ell$ corresponding to the largest eigenvalues. When p, the number of features, is very large as in microarray data, however, it is computationally demanding to directly perform the eigenvalue decomposition (EVD) of $S_g - 2S_\ell$, which is of size $p \times p$. In [26], two approaches for computing LLDA have been presented. The first one directly computes the eigenvalues and eigenvectors, hence demands expensive computational costs. The other approach achieves this via the spectral decomposition of Laplacian matrix, but it still needs to compute the eigenvalues and eigenvectors of a $p \times p$ matrix. Even worse, the eigenvectors corresponding to the non-positive eigenvalues are discarded, thus it does not provide the exact solution to the maximization problem and results in losing discriminatory information.

Here, we propose a novel algorithm for computing W, which is particularly efficient when the feature size is much larger than the sample size, i.e. $p \gg n$, as is often the case with microarray data. The proposed algorithm is based on the following theorem (see the Appendix for the proof).

Theorem 1: Let $P\Lambda Q^T$ be the reduced SVD [11] of $X \in \mathbb{R}^{p \times n}$, where $P \in \mathbb{R}^{p \times n}$ and $Q \in \mathbb{R}^{n \times n}$ are orthonormal matrices and $\Lambda \in \mathbb{R}^{n \times n}$ is a diagonal matrix. Further, let $V \Delta V^T$ be the EVD of a symmetric matrix $\Lambda Q^T (L_g - 2L_\ell) Q \Lambda$, where $V \in \mathbb{R}^{n \times n}$ is an orthonormal matrix and $\Delta \in \mathbb{R}^{n \times n}$ is a diagonal matrix. Then, W is constituted by the eigenvectors in PV corresponding to the largest eigenvalues in Δ.

Note that the main computation of the algorithm consists of the SVD of a $p \times n$ matrix and the EVD of an $n \times n$ matrix. Thus, it is very efficient in the case of $p \gg n$. The previous study [23] first removed the null space of the total scatter matrix via the SVD, thereby reducing the dimensionality of the data to $n - 1$, and then applied the MMC in the reduced space, where the rank of the mean subtracted matrix of X was implicitly assumed to be $n - 1$. However, in a more general case where the samples show multi-colinearity, the rank degenerates to less than $n-1$ and needs to be estimated appropriately. In contrast, the proposed algorithm does not involve the dimension reduction before applying LLDA, allowing to deal with the degenerate case.

In this way, the graph Laplacian representation of the MMC enables both the extension to unsupervised LLDA and the efficiency of the algorithm.

IV. LLDA-RFE: Feature Selection Based on LLDA

The proposed algorithm for LLDA can be used in both supervised and unsupervised cases to extract discriminant features from high-dimensional data often encountered in e.g. face recognition [16], [18], [31], [32], text categorization [5], [32], and microarray cancer classification [18], [32], [33]. In the context of microarray data analysis, the features so extracted correspond to *metagenes*, a linear combination of multiple genes, but we are rather interested in identifying discriminative genes themselves.

To this end, the previous study [23] proposed to combine the MMC with recursive feature elimination (RFE). The MMC-RFE algorithm in [23] recursively removes features with the smallest absolute values of the discriminant vectors of the MMC. The RFE approach has recently proven to be effective with regression [19], [34] as well as with support vector machine (SVM) [13]. In the present study, we propose an unsupervised recursive feature selection method using the discriminant vectors of LLDA to identify features that potentially reveal clusters in the samples.

While the number of discriminant vectors extracted by classical LDA is limited to at most $c-1$, the MMC and LLDA are capable of extracting more than $c-1$ discriminant vectors. It can also be shown that the maximum value of $J_{LLDA}(W)$ with the obtained discriminant vectors is equal to the sum of the corresponding eigenvalues. Because the eigenvalues reflect the discrimination ability, we use only the discriminant vectors corresponding to the positive eigenvalues to calculate the weight of each feature. Let $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_n$ be the eigenvalues in Δ. Then, we define the weight of feature j as the sum of the absolute values of d discriminant vectors in W, i.e. $\sum_{i=1}^{d} \sqrt{\delta_i} |[W]_{ji}|$, where d is the number of positive eigenvalues. Here, the discriminant vectors are weighted by the corresponding eigenvalues.

Our proposed algorithm, LLDA-RFE, can be summarized as follows:

Algorithm: LLDA-RFE

Input: sample matrix : $\boldsymbol{X} \in \mathbb{R}^{p \times n}$
k_ℓ : the number of nearest neighbors
Output: r top-ranked features
0. Set $q \leftarrow p$,
Repeat the following steps until $q = r$
1. Construct the complete and k_ℓ-nearest neighbor graphs on $\boldsymbol{X}$ and compute $\boldsymbol{K}_g, \boldsymbol{K}_\ell, \boldsymbol{L}_g$ and $\boldsymbol{L}_\ell$;
2. Perform the SVD of $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^T$;
3. Compute $\boldsymbol{Z} = \boldsymbol{\Lambda}\boldsymbol{Q}^T(\boldsymbol{L}_g - 2\boldsymbol{L}_\ell)\boldsymbol{Q}\boldsymbol{\Lambda}$;
4. Perform the EVD of $\boldsymbol{Z}$ as $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Delta}\boldsymbol{V}^T$;
5. Set $\boldsymbol{W}$ to the eigenvectors in $\boldsymbol{P}\boldsymbol{V}$ corresponding to the positive eigenvalues in $\boldsymbol{\Delta}$;
6. Remove the jth feature with the smallest weight of $\sum_{i=1}^{d} \sqrt{\delta_i}|[\boldsymbol{W}]_{ji}|$.
7. Set $q \leftarrow q - 1$, form $\boldsymbol{X}$ and go to step 1.

V. Related work on unsupervised feature selection

Data variance is one of the most common unsupervised feature selection criteria, and often used as a baseline method for comparison [15], [28]. Although variance ranking can be useful for selecting features that show large variation across all samples, it is not suited for selecting ones that contribute to characterize different clusters in the samples. Hastie *et al.* [14] developed gene shaving to select informative genes from microarray data. Gene shaving iteratively removes genes having lowest correlation with the leading principal component. Because the principal components are found so that they capture the directions of maximum variance in the data, gene shaving is also unsuitable for identifying genes that reveal different clusters. The assumption that discriminative genes exhibit large variance is not necessarily valid particularly for noisy microarray data, due to the large number of irrelevant genes. Indeed, recent studies [28], [30] have shown that variance ranking, principal component analysis and gene shaving are not effective for yielding distinctive patterns between different classes of samples.

The latest and probably more effective unsupervised methods include Laplacian score [15], the $Q-\alpha$ algorithm [30] and SVD-entropy [28]. Among these, Laplacian score and SVD-entropy are employed for comparison in this study. In the following, we give a brief overview of these two methods.

A. Laplacian score

The idea of Laplacian score is to evaluate each feature by its locality preserving power, which is similar in spirit to Locality Preserving Projection [16].

Let $\boldsymbol{f}_r = (f_{r1}, \ldots, f_{rn})^T$, $r = 1, \ldots, p$, denote the rth feature for n samples. First, we construct a nearest neighbor graph in the same way as for the LLDA-RFE algorithm. Then, we compute the weight matrix $\boldsymbol{K}$, the diagonal matrix $[D]_{ii} = \sum_j [\boldsymbol{K}]_{ji}$, and the graph Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$. Finally, the Laplacian score L_r of the rth feature is computed as

$$L_r = \frac{\tilde{\boldsymbol{f}}_r^T \boldsymbol{L} \tilde{\boldsymbol{f}}_r}{\tilde{\boldsymbol{f}}_r^T \boldsymbol{D} \tilde{\boldsymbol{f}}_r},$$

where

$$\tilde{\boldsymbol{f}}_r = \boldsymbol{f}_r - \frac{\boldsymbol{f}_r^T \boldsymbol{D} \boldsymbol{e}}{\boldsymbol{e}^T \boldsymbol{D} \boldsymbol{e}} \boldsymbol{e}.$$

It is worth noting that Fisher score [8] can be related to Laplacian score as shown in [15].

TABLE I
Characteristics of the Datasets Used in This Study.

Dataset	# samples	# classes	# genes
Leukemia	38	2	7129
Colon cancer	62	2	2000
Medulloblastoma	60	2	7129
Breast cancer	76	2	4918
Lung adenocarcinoma	86	2	7129
MLL	57	3	12582
SRBCT	63	4	2308

B. SVD-entropy

Let us assume that $p > n$ for a given sample matrix $\boldsymbol{X} \in \mathbb{R}^{p \times n}$. Denoting by s_j the singular values of $\boldsymbol{X}$, an SVD-based entropy is defined as [2]:

$$E = -\frac{1}{\log(n_r)} \sum_{j=1}^{n_r} V_j \log(V_j),$$

where

$$V_j = s_j^2 \Big/ \sum_{k=1}^{n_r} s_k^2.$$

Here, $n_r \leq n$ is the number of positive singular values, which is equal to the rank of $\boldsymbol{X}$. Then, the contribution of the ith feature to the entropy is defined as [28]:

$$CE_i = E(\boldsymbol{X}_{[p \times n]}) - E(\boldsymbol{X}_{[(p-1) \times n]}),$$

where $\boldsymbol{X}_{[(p-1)\times n]}$ denotes the sample matrix with the ith feature being removed.

Varshavsky *et al.* [28] have proposed three feature selection strategies based on SVD-entropy: simple ranking (SR), forward selection (FS) and backward elimination (BE). SVD-entropy-based BE has high computational complexity in the case of a large number of features, hence impractical to apply to microarray datasets. This is due to the fact that CE_i is calculated on a leave-one-out basis. Indeed, they used only SR in their experiments on microarrays. Accordingly, we employ SR in this study; top-ranked features are those with the largest values of CE_i.

VI. Experimental results

A. Datasets and preprocessing

In the experiments, we used seven public datasets of cancer microarrays. Since binary classification is a typical and fundamental issue in diagnostic and prognostic prediction of cancer, the different methods were primarily compared using binary-class datasets: ALL versus AML for Leukemia [10], normal versus tumor for Colon cancer [1], outcome prediction on Medulloblastoma [25], Breast cancer [27], and Lung adenocarcinoma [4]. In addition, we used multi-class datasets on MLL [3] and SRBCT [17] to assess their performances. The characteristics of these datasets are summarized in Table I, and the details are given below:

- Leukemia [10]: This Affymetrix high-density oligonucleotide array dataset contains 38 samples from 2 classes of leukemia: 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML). The dataset is publicly available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

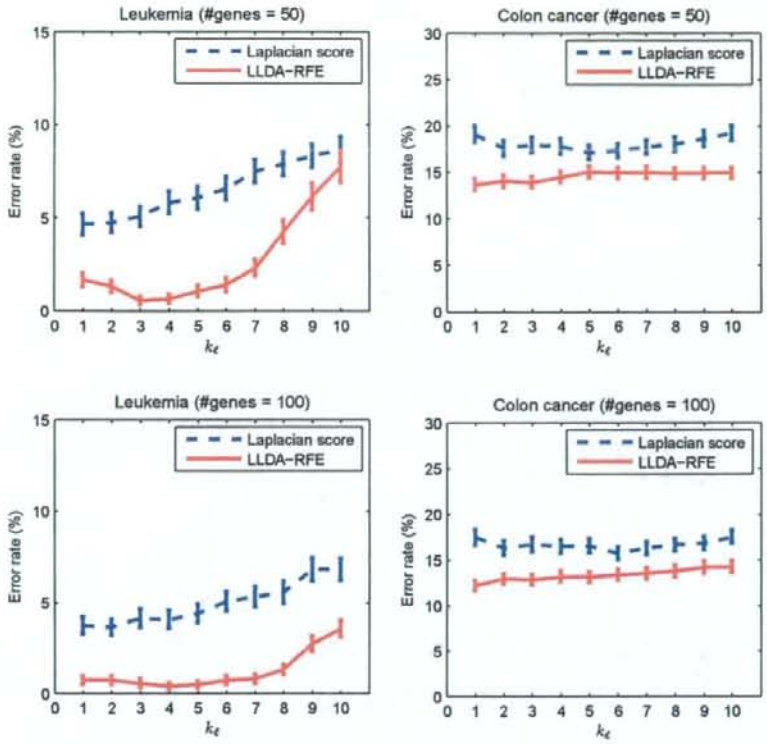

Fig. 1. Comparison between Laplacian score and LLDA-RFE using varying values of k_ℓ for Leukemia and Colon cancer

- Colon cancer [1]: This Affymetrix high-density oligonucleotide array dataset contains 62 samples from 2 classes of colon-cancer patients: 40 normal healthy samples and 22 tumor samples. The dataset is publicly available at http://microarray.princeton.edu/oncology/affydata/index.html.
- Medulloblastoma dataset [25]: This Affymetrix high-density oligonucleotide array dataset contains 60 samples from 2 classes on patient survival with medulloblastoma: 21 treatment failures and 39 survivors. The dataset is publicly available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.
- Breast cancer [27]: This cDNA microarray dataset contains 76 samples from 2 classes on five-year metastasis-free survival: 33 poor prognosis and 43 good prognosis. The dataset is publicly available at http://www.rii.com/publications/2002/vantveer.html.
- Lung adenocarcinoma [4]: This Affymetrix high-density oligonucleotide array dataset contains 86 samples from 2 classes on survival: an event of death for 34 and alive for 52. The dataset is publicly available at http://dot.ped.med.umich.edu:2000/ourimage/pub/Lung/index.html.
- MLL [3]: This Affymetrix high-density oligonucleotide array dataset contains 57 samples from 3 classes of leukemia: 20 acute lymphoblastic leukemia (ALL), 17 mixed-lineage leukemia (MLL), 20 acute myelogenous leukemia (AML). The dataset is publicly available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.
- SRBCT [17]: This cDNA microarray dataset contains 63 samples from 4 classes of small round blue-cell tumors of childhood (SRBCT): 23 Ewing family of tumors, 20 rhabdomyosarcoma, 12 neuroblastoma, and 8 non-Hodgkin lymphoma. The dataset is publicly available at http://research.nhgri.nih.gov/microarray/Supplement/.

For the Leukemia, Medulloblastoma, Lung adenocarcinoma and MLL datasets, expression values were first thresholded with a floor of 100 and a ceiling of 16000, followed by a base 10 logarithmic transform. Then, each sample was standardized to zero mean and unit variance across genes. For the Colon cancer dataset, after a base 10 logarithmic transform, each sample was standardized. For the Breast cancer dataset, after the filtering of genes following [27], each sample was standardized. For the SRBCT dataset, the expression profiles already preprocessed following [17] were used.

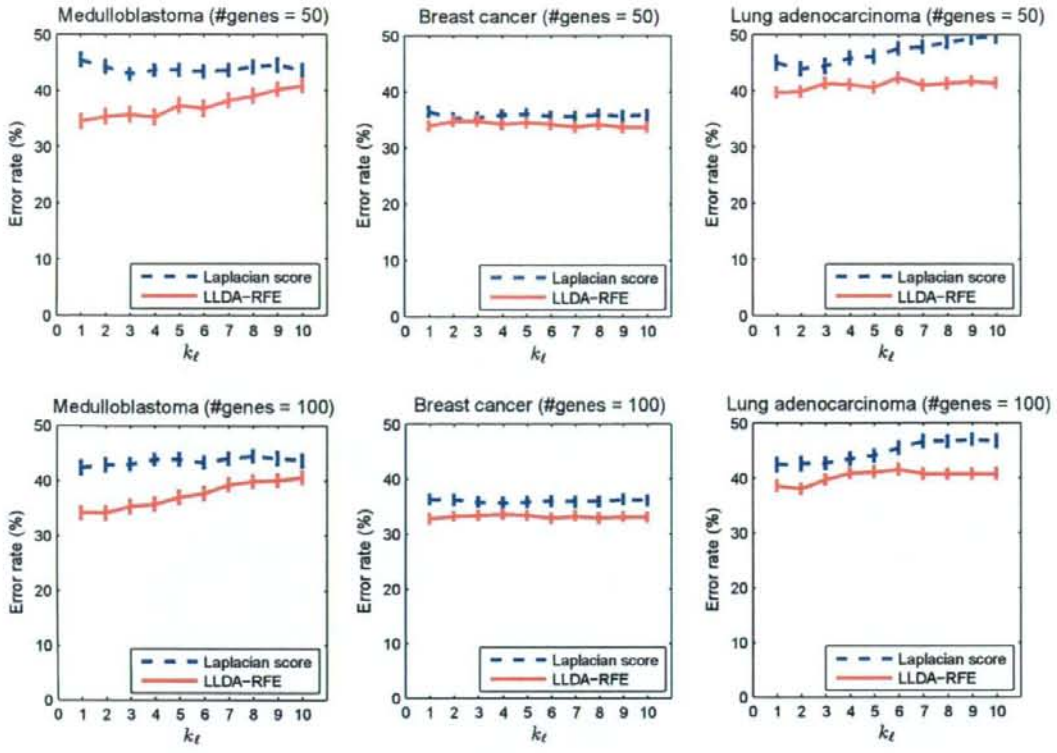

Fig. 2. Comparison between Laplacian score and LLDA-RFE using varying values of k_ℓ for Medulloblastoma, Breast cancer and Lung adenocarcinoma

B. Performance evaluation and experimental settings

We compare the performance of LLDA-RFE with those of state-of-the-art unsupervised feature selection methods, Laplacian score and SVD-entropy. The performances of the unsupervised methods are evaluated by their capability of identifying discriminative genes without using class information. Varshavsky *et al.* [28] employed the Jaccard score of clustering algorithms such as K-means, showing how clusters can be discovered by using a smaller number of genes selected from several thousand or more genes in the same samples. Wolf and Shashua [30] measured the performances by the classification accuracy of a linear SVM classifier using leave-one-out cross-validation; gene selection was performed in an unsupervised setting, but classification in a supervised setting using selected genes only. Because we also compare the performance between LLDA-RFE and a supervised gene selection method, Fisher score [8], we employed the nearest mean classifier (NMC) and measured the performances by its classification accuracy. It is known that NMC is highly effective for cancer classification despite its simplicity [29]. Note that since Fisher score is a supervised filter, it is generally expected to perform better than unsupervised methods.

We assessed the performance of each gene selection method with NMC by repeated random splitting as in [22]; the samples were partitioned randomly in a class proportional manner into a training set consisting of two-thirds of the whole samples and a test set consisting of the held-out one-third of the samples. To avoid selection bias, gene selection was performed using only the training set, and the classification error rate of the learnt classifier was obtained using the test set. This splitting was repeated 100 times. The error rates averaged over the 100 trials and the corresponding standard error rates are reported here.

To save computational time of RFE, we removed half of the genes until less than 500, and then a single gene at a time. For the computation of graph Laplacians, we used the Euclidean distance for nearest neighbor search and a simple 0-1 weighting as the similarity of the graph, i.e. $k(x_i, x_j) = 1$ if x_i and x_j are connected, and 0 otherwise.

C. Results and discussion

1) Effect of k_ℓ: We first compare the performance between LLDA-RFE and Laplacian score by varying the number of nearest neighbors, on the binary-class datasets: Leukemia, Colon cancer, Medulloblastoma, Breast cancer and Lung adenocarcinoma. Figs. 1 and 2 show the average error and standard error rates for $k_\ell = 1, \ldots, 10$. For LLDA-RFE, k_ℓ was fixed to the same value

during elimination. The number of genes selected and used for classification is 50 and 100.

It is clear that LLDA-RFE consistently achieves better performance than Laplacian score. This can be attributed to the difference that while Laplacian score is univariate, LLDA-RFE is multivariate and gene subsets are refined by the recursive elimination.

It may be difficult to set an appropriate value of k_ℓ in fully unsupervised settings, because we cannot rely on cross-validation unless class labels are provided and the value can also be largely dependent on the sample size of each dataset and on the potential number of clusters therein. Although an adaptive setting of the value might be preferable during elimination, our results suggest that $k_\ell = 1-3$ is a reasonable choice when applying LLDA-RFE to microarray datasets with small sample size.

2) Comparison on binary-class datasets: Table II shows the average error and standard error rates of NMC with four gene selection methods on the binary-class datasets. Fig. 3 plots the average error rates as a function of the number of genes from 1 to 100. The number of nearest neighbors for Laplacian score was set as follows: $k_\ell = 2$ for Leukemia, $k_\ell = 6$ for Colon cancer, $k_\ell = 3$ for Medulloblastoma and Breast cancer, and $k_\ell = 1$ for Lung adenocarcinoma. For LLDA-RFE, $k_\ell = 3$ was used for Leukemia and $k_\ell = 1$ for the other datasets.

We can observe that LLDA-RFE outperforms Laplacian score for a wide range of gene sizes. In comparison with SVD-entropy, LLDA-RFE yields lower error rates for Leukemia, Medulloblastoma and Breast cancer. Although SVD-entropy performs better for Colon cancer and Lung adenocarcinoma, LLDA-RFE consistently shows satisfactory performances for all the datasets. Also, note that LLDA-RFE performs better than Fisher score for Leukemia, Medulloblastoma and Breast cancer, despite the fact that LLDA-RFE is fully unsupervised. However, this does not imply that unsupervised gene selection is preferred to supervised one for these datasets. In fact, Fisher score, which can be viewed as a supervised version of Laplacian score, improves the performance of Laplacian score by using class information. Likewise, we can expect further improvement when using LLDA-RFE in a supervised manner.

3) Comparison on multi-class datasets: Table III shows the average error and standard error rates for the MLL and SRBCT datasets. Fig. 4 plots the average error rates as a function of the number of genes from 1 to 100. For Laplacian score, $k_\ell = 1$ was used for both datasets, and for LLDA-RFE, $k_\ell = 4$ and 3 were used for MLL and SRBCT, respectively.

It can be seen that LLDA-RFE reaches smaller error rates with a smaller number of genes, showing superior performance to Laplacian score and SVD-entropy. Notably, LLDA-RFE achieves even better performance than Fisher score. These results indicate that LLDA-RFE can also be useful for filtering genes from microarray samples potentially comprising multiple clusters.

In summary, our comparison using several microarray datasets has demonstrated that LLDA-RFE is effective for identifying genes that contribute to characterize different clusters in the samples. Although we used 0-1 weighting as the similarity measure, the performance could be improved by using other data-dependent similarity measures. Also, more discriminative features can be found by balancing the global and local scatters as in (3).

TABLE II
COMPARISON ON BINARY-CLASS DATASETS. BEST RESULTS IN BOLD FACE.

# genes	Fisher score	SVD-entropy	Laplacian score	LLDA-RFE
Leukemia				
20	4.9 ± 0.6	**2.6 ± 0.4**	9.6 ± 1.0	3.6 ± 0.5
50	3.9 ± 0.4	1.6 ± 0.4	4.8 ± 0.5	**0.6 ± 0.2**
100	2.9 ± 0.4	1.6 ± 0.4	3.7 ± 0.4	**0.6 ± 0.2**
Colon cancer				
20	**12.4 ± 0.6**	14.9 ± 0.6	18.2 ± 0.8	15.9 ± 0.6
50	13.0 ± 0.6	**11.5 ± 0.6**	17.4 ± 0.7	13.7 ± 0.6
100	12.8 ± 0.5	**11.7 ± 0.6**	15.7 ± 0.7	12.2 ± 0.6
Medulloblastoma				
20	38.8 ± 0.9	36.8 ± 1.1	43.7 ± 1.0	**34.1 ± 1.1**
50	38.7 ± 1.0	38.4 ± 1.0	43.0 ± 0.9	**34.6 ± 1.1**
100	38.5 ± 1.0	37.1 ± 1.0	43.1 ± 1.0	**34.2 ± 1.1**
Breast cancer				
20	35.2 ± 0.9	42.6 ± 0.9	35.1 ± 0.8	**33.3 ± 0.8**
50	36.0 ± 0.8	42.0 ± 0.8	35.4 ± 0.8	**33.8 ± 0.7**
100	36.3 ± 0.8	42.4 ± 0.8	35.9 ± 0.7	**32.8 ± 0.7**
Lung adenocarcinoma				
20	**37.8 ± 0.8**	40.5 ± 0.8	45.7 ± 1.2	42.6 ± 0.7
50	**36.3 ± 0.8**	40.0 ± 0.7	45.0 ± 1.1	39.7 ± 0.8
100	**35.1 ± 0.8**	38.3 ± 0.8	42.4 ± 1.1	38.6 ± 0.8

TABLE III
COMPARISON ON MULTI-CLASS DATASETS. BEST RESULTS IN BOLD FACE.

# genes	Fisher score	SVD-entropy	Laplacian score	LLDA-RFE
MLL				
20	7.2 ± 0.5	26.9 ± 0.9	10.2 ± 0.8	**6.1 ± 0.5**
50	6.6 ± 0.5	8.1 ± 0.6	9.4 ± 0.6	**5.1 ± 0.5**
100	5.9 ± 0.5	4.8 ± 0.4	9.1 ± 0.6	**3.8 ± 0.4**
SRBCT				
20	**3.6 ± 0.5**	22.6 ± 1.0	17.4 ± 1.2	16.2 ± 1.0
50	**2.6 ± 0.4**	17.4 ± 0.8	13.4 ± 1.0	12.0 ± 0.7
100	**4.6 ± 0.4**	11.8 ± 0.7	11.3 ± 0.9	11.4 ± 0.7

VII. CONCLUSIONS

In this paper, we have proposed a new unsupervised feature selection method based on Laplacian linear discriminant analysis (LLDA). In particular, we have extended LLDA to unsupervised cases and proposed an efficient algorithm for computing the discriminant vectors of LLDA. The LLDA-based Recursive Feature Elimination (LLDA-RFE) algorithm was applied to several microarray datasets to identify discriminative genes without using class labels.

Our comparison with other state-of-the-art unsupervised feature selection methods and with a supervised filter method has demonstrated the feasibility and effectiveness of the proposed algorithm. LLDA-RFE is capable of identifying discriminative features that contribute to reveal underlying class structures, providing a useful tool for the exploratory analysis of biological data.

A possible application of interest is the use of LLDA-RFE in semi-supervised scenarios; when labels are partially given, we