Analyse molekularbiologischer Daten mittels Self-organizing Maps

Dr. rer. nat. Henry Wirth

Interdisziplinäres Zentrum für Bioinformatik, Universität Leipzig
LIFE Forschungszentrum für Zivilisationserkrankungen, Universität Leipzig
Helmholtz-Zentrum für Umweltforschung, Leipzig
wirth@izbi.uni-leipzig.de

Abstract: Die Molekularbiologie sieht sich gegenwärtig mit enormen Datenmengen konfrontiert, welche durch moderne Hochdurchsatzmessungen wie Microarrays oder Sequenzierung erzeugt werden. Etablierte Methoden zur Datenanalyse erweisen sich zumeist als ungeeignet für solch hochdimensionale, komplexe und meist verrauschte Daten.
Wir stellen hier kurz unsere Analysestrategie vor, welche essenzielle Methoden wie Dimensionsreduzierung, Clustering, multidimensionales Skalieren und Visualisierung vereint und zudem eine in hohem Maße intuitive Sicht auf die Daten erlaubt.

1 Einführung

Die moderne Molekularbiologie sieht sich gegenwärtig mit enormen Datenmengen konfrontiert. Immer geringer werdende Kosten für Hochdurchsatz-Analysen wie Microarrays, Sequenzierung oder Massenspektrometrie führten in den letzten Jahren zu einer Explosion der Menge an produzierten und in Datenbanken verfügbaren molekularbiologischen Daten. Die erhobenen Datensätze umfassen dabei oft mehrere hundert Proben mit mehreren zehntausend, teilweise millionen gemessenen Merkmalen (sog. Features) je Probe. Für die bioinformatische Auswertung sind solche Datenfluten eine große Herausforderung und erfordern optimale Analysestrategien.
Besonders wichtige aber auch komplizierte Aufgabenstellungen wie die Vorverarbeitung der Rohdaten, ihre passende Evaluierung und Wichtung entsprechend ihres Einflusses auf die Biologie und geeignete Methoden zur Auswahl und Interpretation relevanter Features beschäftigen die moderne Bioanalytik. Viele weit verbreitete Methoden wie Cluster-Heatmaps, Balken- oder Streu-Diagramme geben schnell und intuitiv Überblick über die Daten. Andererseits repräsentieren viele der etablierten und einfachen Ansätze den multivariaten Charakter der Daten nur ungenügend, sodass für das Verständnis des untersuchten Systems wichtige Informationen verborgen bleiben können. Mehr noch, die moderne Medizin und Zellbiologie verlangt die individuelle Analyse jeder Probe. Dies erfordert ein Portraitieren der Datenlandschaft jedes einzelnen Patienten oder jeder einzelnen Zelle und das Ausarbeiten individueller Unterschiede beispielsweise in Folge verschiedener Behandlungen oder von Entwicklungsprozessen.

Geeignete Analysestrategien umfassen in diesem Zusammenhang aufeinander abgestimmte Visualisierungen der Messwerte und der Ergebnisse. Die beiden Sichten auf die Daten, entweder Proben- oder Feature-basiert, sollten in geeigneter Weise kombiniert werden und so die Besonderheiten jeder einzelnen Probe in ihrem multivariaten Kontext zugänglich machen. Diese Selektion sogenannter Biomarker erfordert die Anwendung passender Statistiken zur Schätzung der Signifikanz sowie folgender Analysemethoden um den funktionellen Zusammenhang zu erschließen.
Einen besonderen Ansatz für die genannten Aufgaben stellen die Methoden des maschinellen Lernens dar. Besonders neuronale Netze wie die selbstorganisierenden Karten (engl. self-organizing maps, SOMs) kombinieren effektive Verarbeitung und Reduktion großer Datenmengen mit einzigartigen Visualisierungsmöglichkeiten. SOMs bieten daher ein passendes Herzstück um große und komplexe Datensätze detailliert zu untersuchen. Die vorliegende Arbeit soll nun unsere Analysestrategie auf Basis SOM-verarbeiteter molekularbiologischer Daten kurz zusammenfassen und ihre Anwendung auf verschiedene experimentelle Systeme sowie Messmethoden illustrieren.

2 SOM Portraitierung und Datenkompression

Das von uns entworfene und implementierte Softwarepaket umfasst zehn Module, welche jeweils individuell angepasste und integrierte Methoden enthalten. Die Module können grob in drei Abschnitte eingeteilt werden (siehe Abbildung 1, rechte Seite): Verarbeitung der Einzel-Features, Analysen auf Basis der Meta-Features und Analysen der Fleckenmuster der Proben-Portraits. Die Vorverarbeitung wandelt die Rohdaten in Trainingsdaten für das maschinelle Lernen um. Dabei sollen technisch bedingte Verzerrungen und Artefakte entfernt und die Vergleichbarkeit der Proben gewährleistet werden. Je nach Messmethode werden dabei unterschiedliche Kalibrierungs- und Normalisierungsalgorithmen angewendet.
Die vorverarbeiteten Daten liegen dann als Feature-Matrix der Dimension N x M vor (siehe Abbildung 1a), wobei N die Anzahl der Features bezeichnet, die je Probe gemessen werden, und M die Anzahl der Proben (z.B. Patienten, unterschiedliche Behandlungen oder verschiedene Zeitpunkte) im Datensatz. Dabei ist N typischerweise um Größenordnungen größer als M. Die Zeilen der Datenmatrix werden als Profile der Features bezeichnet, die Spalten als Zustder Proben. Ziel des SOM Lernens ist nun die Reduktion der Zahl der relevanten Features. Dies wird durch eine Sortierung und Gruppierung ähnlicher Feature-Profile erreicht. Dabei wir die Eingangs-Datenmatrix in eine sogenannte Meta-Datenmatrix überführt welche eine wesentlich geringere Anzahl an Meta-Features enthält (Abbildung 1b). Anders ausgedrückt projiziert die SOM den hochdimensionalen Eingangsdatenraum auf einen Meta-Datenraum geringerer Dimension. Dadurch wird der Datenraum in Gruppen ähnlicher Feature-Profile segmentiert, welche zu je einem Meta-Feature gehören. Dieses Feature-Mapping ist in Abbildung 1a und b durch rote Linien dargestellt. Die von uns verwendete Standard-SOM verwendet dazu K in einem zweidimensionalen Raster angeordnete Knoten, hinter jedem sich ein Meta-Feature repräsentiert durch ein Profil der Länge M verbirgt.

Die Spalten der Meta-Datenmatrix, die Meta-Zustände der Proben, können durch geeignete Färbung der entsprechenden Werte (rot=hohe Werte, grün=mittlere Werte, blau=niedrige Werte) als intuitive Mosaikbilder mit charakteristischen Fleckenmustern visualisiert werden (siehe auch Abbildungen 2-4). Dabei ist besonders bemerkenswert, dass diese zweidimensionalen Portraits die multivariate Datenlandschaft detailgetreu widergeben und daher alternative Visualisierungsmethoden wie die beliebten, jedoch univariaten Heatmaps ausstechen. Der SOM Algorithmus assoziiert ähnliche Feature-Profile, welche zum gleichen Hauptmodul der Variabilität gehören, zum gleichen beziehungsweise zu benachbarten Meta-Features, wogegen unabhängige Features zu entfernteren Meta-Features assoziiert werden. Durch diese Eigenschaft des SOM Algorithmus ergeben sich intuitive Fleckenmuster mit gleichmäßigen Farbverläufen und roten bzw. blauen fleckenartigen Regionen, welche jeweils Gruppen von Features mit besonders hohen bzw. niedrigen Werten repräsentieren (Abbildung 1c). Meta-Features, und damit auch alle assoziierte Einzelfeatures, können als koreguliert betrachtet werden, wogegen separierte Flecken verschiedene regulatorische Moden repräsentieren. Dieser globale Blick auf die Feature-Landschaft spiegelt die multivariate Kovarianzstruktur der Daten wieder und ist wesentlich intuitiver als das Suchen nach Unterschieden zwischen den Proben in sortierten Listen hunderter Einzelfeatures, wie sie Standardanalysen produzieren. Daher können die Flecken-Module als natürliche Wahl angesehen werden wenn es darum geht, kontextbezogene Muster in komplexen Datensätzen zu identifizieren.
Es ist wichtig zu erwähnen, dass dieses Zerlegen der Meta-Features in koregulierte Module neben dem SOM Lernen eine zweite Ebene der Datenkompression darstellt. Beispielsweise werden die mehr als 20,000 Gene eines Transkriptom-Datensatzes mit knapp 70 Proben menschlicher Gewebe durch den SOM Algorithmus auf 60 x 60 = 3,600 Meta-Gene reduziert [WLvBB11]. In den entsprechenden Portraits ließen sich ein Dutzend unterschiedliche Flecken-Module identifizieren. Somit wurde die Datenmenge um etwa drei Größenordnungen reduziert, ohne dabei in irgendeiner Weise Primärinformationen zu verlieren. Denn anders als bei Filter-Methoden wird dies durch eine Wichtung und Aggregation der gesamten Information erreicht. Sie ist auch nach der Transformation noch vollständig in den Meta-Daten enthalten und kann in nachgeschalteten Analysen wieder abgerufen werden.

3 Information Mining

Die Erste der Analysen auf Basis der Meta-Daten zielt auf das Identifizieren der Module potentiell koregulierter Features ab (Abbildung 1c). Wir verwenden dazu verschiedene Metriken: Erstens können Regionen mit besonders hohen bzw. niedrigen Werten (also besonders rote bzw. blaue Meta-Features) durch Anwenden eines einfachen Perzentil-Kriteriums identifiziert werden. Zweitens können etablierte Clusterverfahren wie zum Beispiel k-Means angewendet werden, um den durch die SOM aufgespannten Raum zu segmentieren. Zusätzliche Supporting Maps visualisieren verschiedene wichtige Eigenschaften der SOM und der Meta-Datenlandschaft, wie etwa Überblickskarten welche sämtliche identifizierten Flecken enthalten, eine Populationskarte welche die Anzahl der assoziier-

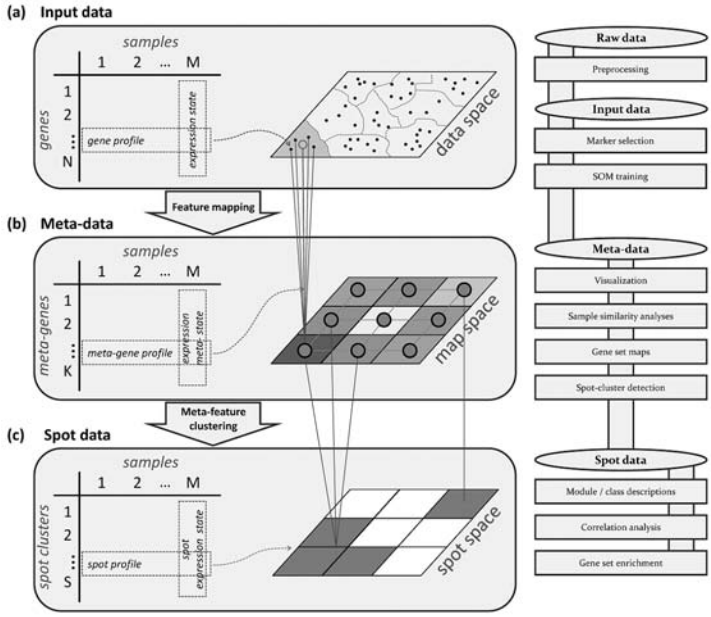

Abbildung 1: Zweistufige Datenkompression bei Anwendung des SOM Lernens: Zunächst werden die Eingangsdaten in Meta-Daten transformiert, danach werden die Meta-Daten anhand der charakteristischen Fleckenmuster gruppiert. Auf der rechten Seite ist der Ablaufplan der gesamten Analyse illustriert.

ten Einzelfeatures pro Meta-Feature aufzeigt, oder Varianz- und Entropiekarten der Meta-Feature Profile. Information Mining im engeren Sinne beinhaltet Fragestellungen wie Biomarker Selektion, das Klassifizieren der Proben und das Einbeziehen funktioneller Analysen. Die SOM Methodik erlaubt es uns nun solche Aufgaben auf den komprimierten Daten auszuführen. Wir konnten zeigen, dass dies im Vergleich zu Analysen auf Basis der unkomprimierten Daten nicht nur Vorteile bei Rechenzeit- und Speicherplatzbedarf bietet, sondern auch die Ergebnisse in Bezug auf Sensitivität und Spezifität verbessert [WLvBB11]. Daher wurde eine Reihe etablierter statistischer Methoden für die Anwendung auf die Meta-Daten angepasst und implementiert um Aufgaben wie Marker Selektion, Ähnlichkeits-und Klassenanalysen der Proben oder funktionelle Analysen der differentiellen Muster zu realisieren [WvBB12].

4 Anwendungsbeispiele

In diesem Abschnitt soll die SOM Analyse anhand dreier ausgewählter Anwendungsbeispiele illustriert werden. Diese wurden stellvertretend für verschiedene 'OMe' (Transkriptom, Genom, Proteom) und verschiedene experimentelle Designs (Entwicklungsserien,

weltweite Populationskohorte und Proben unterschiedlicher Taxa) ausgewählt.

4.1 Transkriptomdynamik des Hefezyklus aus Sicht der SOM Expressionsportraits

Der Stoffwechselzyklus der Backhefe (engl. 'yeast metabolic cycle', YMC) ist eines der am besten untersuchten Modelle der molekularbiologischen Forschung. Wir nutzen dieses Wissen um unsere SOM Methodik in Bezug auf Dynamiken der Genexpression zu evaluieren. Während des etwa 45minütigen Zyklus' wurden die Expressionslevel von knapp 6,000 Genen der Hefe in Zeitintervallen von 4 Minuten gemessen [LK06]. Jede der 11 Proben (t1 - t11) wurde in ein Expressionsportrait transformiert, welches rote und blaue Flecken entsprechend über- und unterexprimierten (Meta-)Genen enthält (siehe Abbildung 2a). Diese Flecken scheinen gegen den Uhrzeigersinn entlang der Ränder der Portraits zu rotieren und spiegeln somit den gleitenden Übergang der transkriptionellen Zustände zwischen den jeweiligen Zeitpunkten wieder. Die Heatmap in Abbildung 2b zeigt den Expressionsverlauf sechs identifizierter Flecken der Portraits (siehe die Bilder auf der linken Seite). Jeder der Flecken enthält konzertierte Meta-Gene, bzw. mit ihnen assoziierte Einzelgene, welche mit größer werdender Phasenverschiebung oszillieren. Die Analyse angereicherter Gensets erlaubt es uns nun jedem dieser regulatorischen Modulen sein transkriptionelles Programm im Kontext der Zellfunktionen zuzuordnen (Abbildung 2c). Interessanterweise liefert die SOM Portraitierung nicht nur mit den drei Hauptphasen des Zyklus' konsistente Ergebnisse [TKRM05]. Sie bietet auch eine verfeinerte Sichtweise auf die Dynamik des Transkriptoms. Die Expressionsprofile ausgewählter funktioneller Gensets sind in Abbildung 2d zu sehen. Auch auf dieser funktionellen Ebene ist das zyklisch oszillierende Verhalten sowie die phasenverschobene Zeitabhängigkeit der Genexpression detailliert analysierbar.

4.2 Genomportraits enthüllen Fußspuren der frühmenschlichen Völkerwanderung

Die genetische Diversität der Menschheit wird durch biologische als auch durch demografische Faktoren geformt und ihr Verständnis ist fundamental für die detaillierte Untersuchung der genetischen Grundlagen verschiedenster Krankheiten. In dieser Anwendung haben wir die SNP (= 'single nucleotide polymorphism') Landschaften von über 1,000 Menschen aus 57 ethnischen Gruppierungen und 7 geographischen Regionen untersucht. Die Daten wurden durch das 'Human Genome Diversity Project' erhoben und beinhalten Informationen über mehr als 650,000 SNPs pro Individuum [LAT$^+$08]. Unser SOM Ansatz komprimiert diesen riesigen Datensatz mit erstaunlich hoher Auslösung auf 80 x80 = 6,400 'Meta-Allele'. Um den Datensatz für maschinelles Lernen verarbeitbar zu machen wurde ein ternärer Code verwendet: '0' für das häufigere und '2' für das seltenere homozygote Allel (blaue bzw. rote Färbung in den SOM Portraits), '1' für heterozygote Allele (grüne Färbung). Die von der SOM erzeugten Portraits jedes Individuums wurden entsprechend der geografischen Zugehörigkeit gemittelt (siehe Abbildung 3). Diese Regionen-Portraits spiegeln den lokalen SNP Mutationsstand in ihren charakteristischen Fleckenmustern wieder. Dabei divergiert die Ähnlichkeit der Portraits mit zunehmender

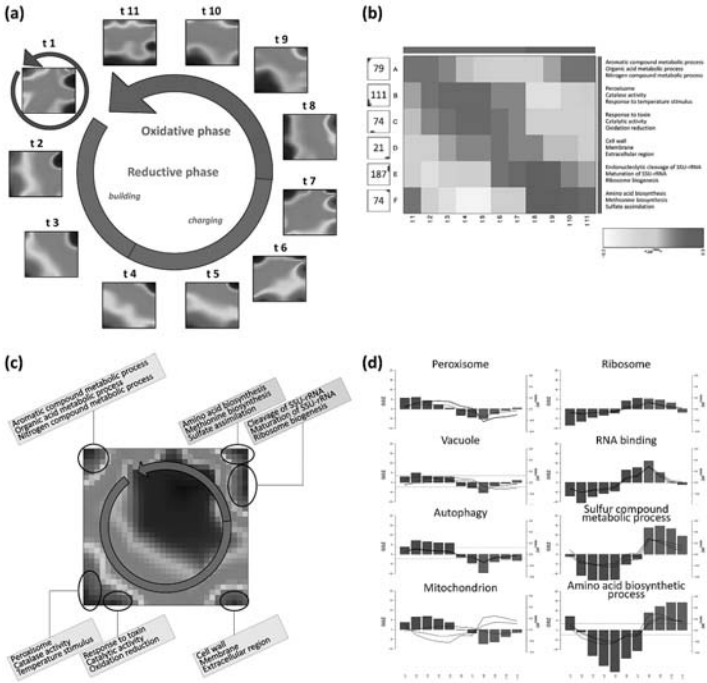

Abbildung 2: Portraits der Expressionsdynamik des Hefezyklus: (a) Die SOM Portraits der Proben spiegeln den zyklischen Charakter der Expressionslandschaften wieder. (b) Sechs verschiedene rote Flecken wurden in den Portraits identifiziert. Die Heatmap stellt deren Profile über die Zeitpunkte dar. Diese oszillieren, weisen jedoch verschiedene Zeitpunkte der maximalen Amplitude auf. (c) Die Überblickskarte sammelt alle überexprimierten (roten) Flecken und ordnet ihnen angereicherte funktionelle Gensets zu. (d) Die Expressionsprofile ausgewählter Gensets zeigen den Zeitverlauf verschiedener transkriptioneller Programme.

geografischer Entfernung der Populationen vom vermuteten Ursprung der Menschheit in Afrika. Diese Divergenz erfolgt jedoch nicht abrupt sondern als gleichmäßiger Übergang, was die zeitliche Veränderung der Mutations-Muster durch Migration einerseits und die genetische Isolation geografisch getrennter Populationen andererseits widerspiegelt. Bemerkenswerterweise durchbricht das Portrait von Ozeanien die Ähnlichkeitsregel benachbarter Regionen. Es zeigt vielmehr ein gesprenkeltes Muster mit vielen kleinen Flecken entlang der Ränder des Portraits. Die Einwohner Ozeaniens teilen sich demnach genetische Charakteristiken mit nahezu allen anderen Regionen. Dieses spezifische Muster deutet auf eine frühere Migrationswelle der Frühmenschen zusätzlich zur 'Out-of-Africa' Wanderung hin. Eine aktuelle Studie zeigte kürzlich, dass Melanesier und Papua im Vergleich zu den anderen nicht-afrikanischen Populationen genetische Ähnlichkeit mit den Neandertalern und den Denisova-Menschen aufweisen [RGK$^+$10]. Die wurde damit erklärt, dass die Frühmenschen möglicherweise in mehreren Wellen migrierten. Um die verschiedenartigen Fleckenmuster der ozeanischen Individuen zu erklären muss man sich verdeut-

lichen, dass die meisten anderen Populationen aus einer gemeinsamen Migrationswelle stammen. Diese große 'Kohorte' mit ihren überlappenden, weil kontinuierlich evolvierten SNP-Landschaften hat auch einen großen Anteil an der Formung des Meta-Datenraumes. Populationen einer anderen Migrationswelle, wie die Papua-Melanesier, passen in diese sequentiellen SNP-Zustände erwartungsgemäß in weit geringerem Maße und erzeugen ein stärker fragmentiertes Fleckenmuster. Andererseits führten Vermischungen mit den Genpools der angrenzenden Populationen Südostasiens dazu, dass einige ozeanische Individuen den asiatischen sehr ähnlich sind. Wir haben für die mehr als eintausend untersuchten Menschen einen Korrelations-Spannbaum erzeugt, in welchem Individuen mit der höchsten Korrelation ihrer SNP-Landschaften paarweise verbunden sind (Abbildung 3). Erstaunlicherweise passt dieser Baum sehr genau auf die heutige geografische Verteilung der Populationen. Doch auch interessante Details sind klar erkennbar. Die Afrikanischen Populationen lassen sich genau in zwei Zeige unterteilen: Jäger und Sammler (z.B. San und Pygmy) und Agrargesellschaften (Bantu, Yoruba). Populationen und Individuen aus dem mittleren Osten, Europa und Zentralasien vermischen sich teilweise entsprechend den komplexen Wanderungen im Laufe der Geschichte. In Ostasien bilden die Jakuten eine 'Brücke' zu den amerikanischen Ureinwohnern während die Kambodschaner zu den Populationen Ozeaniens verbunden sind. Diese Ergebnisse untermauern die Feststellung, dass Ähnlichkeiten zwischen den SNP-Landschaften mit zunehmender zeitlicher und räumlicher 'Entkopplung' der Populationen abnehmen [LAT+08]. Somit offenbaren die SOM Portraits und der Korrelations-Baum Fußspuren der frühmenschlichen Migration in unseren Genomen. Dieses Beispiel zeigt die Fähigkeit des SOM Lernens eine große Anzahl komplexer Genotypen zu verarbeiten und für intuitive und dennoch detaillierte Analysen aufzubereiten.

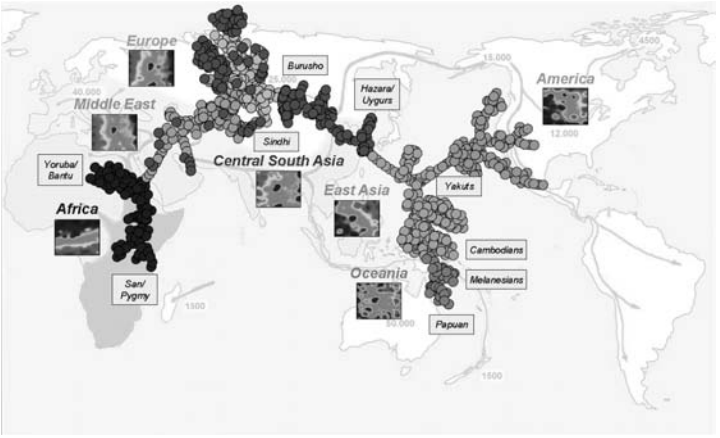

Abbildung 3: SOM Analyse der genetischen Diversität des Menschen: Der Korrelations-Spannbaum visualisiert die Ähnlichkeitsrelationen zwischen den SNP-Landschaften der Individuen. Jeder Kreis entspricht einem Individuum und ist nach geografischer Zugehörigkeit gefärbt. Jede der Regionen ist durch das mittlere SOM Portrait der entsprechenden Individuen repräsentiert und eindeutig charakterisiert. Die Struktur des Baumes spiegelt die gegenwärtige Hypothese der frühmenschlichen Verbreitung von Afrika aus wieder.

4.3 Proteom-Portraits der taxonomischen Verwandtschaftsverhältnisse von Drosophila

Massenspektrometrie (MS) ist neben Microarrays und Sequenzierung eine weitere Technologie, welche im Hochdurchsatzbereich enorme Mengen an Daten produziert und in der Proteomik und Metabolomik weit verbreitet ist. Dabei ist besonders die 'MALDI-ToF MS' Methode zu erwähnen, welche Informationen über Peptide und Proteine einer Probe in einem breiten Massenbereich liefert. Sie wird unter anderem als 'MALDI-typing' angewandt und ermöglicht die Klassifizierung der Proben anhand ihrer gesamten Spektralinformation und ohne die dahinterliegende Proteinzusammensetzung explizit zu kennen. In diesem Beispiel analysierten wir MALDI-ToF-Spektren von 125 Drosophila Exemplaren gehörend zu 13 verschiedenen Spezies [FGK+10]. Sogenannte 'Peak-Listen' von 208 Intensitäts-Amplituden je Probe wurden dafür aus den kontinuierlichen Spektren extrahiert und für das Training einer SOM mit 40 x 40 = 1,600 Meta-Features benutzt. Die meisten der pro Spezies gemittelten SOM Portraits weisen nur einen roten Fleck entsprechend Peaks mit hohen Intensitäten auf (Abbildung 4a). Die Position dieses roten Bereiches in den Portraits ist spezifisch für jede Spezies. Das bedeutet, dass es für jede Spezies charakteristische Bereiche in den Spektren gibt, welche nur in dieser eine hohe Intensitätsamplitude haben. Somit bietet unsere Portraitierung eine einfache und direkte M¨glichkeit diese 'MS-Fingerabdrücke' zu extrahieren. Auf Basis der Meta-Feature Landschaften wurde ein phylogenetischer Clusterbaum generiert (Abbildung 4a). Dieser weist sauber getrennte Äste auf, welche die allgemein anerkannte Phylogenie von Drosophila sehr gut widerspiegelt (vergleiche Abbildung 4b). Im Allgemeinen ist festzuhalten, dass phylogenetische Ähnlichkeit direkt mit Ähnlichkeit der Fleckenmuster in den SOM Portraits einhergeht. Einige der Portraits haben zwei rote Flecken, wobei einer oder beide mit der Charakteristik anderer Spezies Übereinstimmen. Zum Beispiel zeigen die Portraits von D. teissieri zwei Flecken an der Position, welche für D. mauritiana bzw. D. funebris spezifisch sind. Daher sind die D. teissieri Proben im phylogenetischen Baum an den Ästen zu finden, welche hauptsächlich mit letzteren beiden Taxa besetzt sind. Die Ursache dieser Spaltung der Protein-Landschaft ist jedoch unklar. Die Originalpublikation berichtet zwar von Problemen bei der Annotation von Teilen der D. teissieri Proben, geht aber nicht auf Details ein [FGK+10]. Andererseits wird das hohe Auflösungsvermögen der Meta-Feature Landschaften dadurch bestätigt, dass die beiden Spezies mit der geringsten evolutionären Distanz, D. miranda und D. speudoobscura mit weniger als 2 Millionen Jahren, im phylogenetischen Baum klar voneinander getrennt sind. Zu beachten ist, dass die Klassifizierung auf Basis der MS-Spektren als problematisch eingestuft wird [FGK+10]. Wie wir jedoch zeigen konnten erhöht die Transformation der Daten in Meta-Daten das Auflösungsvermögen und die Qualität von Klassifizierungen beim MALDI-Typing [WvBB12]. Somit hat sich unsere SOM Analyse auch in diesem schwierigen Teilgebiet verdient gemacht.

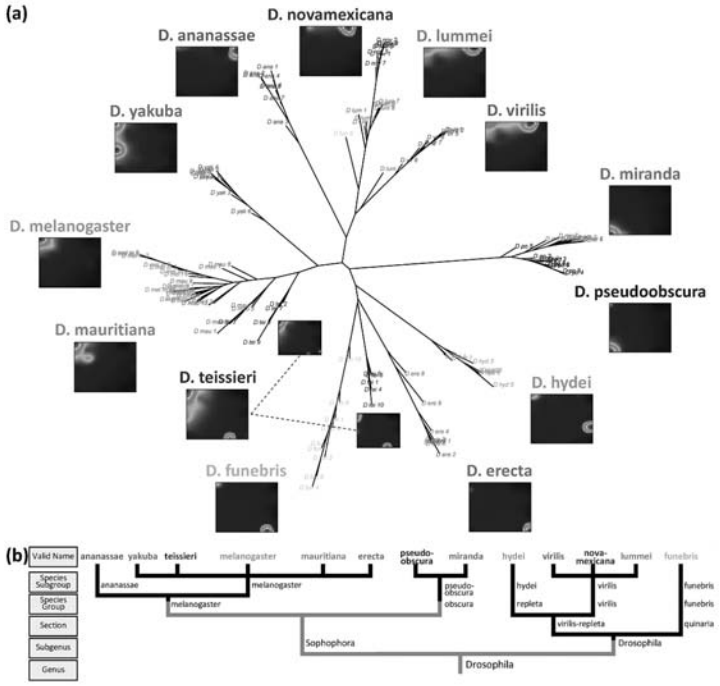

Abbildung 4: Portraitierung des Drosophila-Proteoms: (a) Der Ähnlichkeitsbaum der Drosophila Meta-Feature Landschaften wurde durch den 'Neighbor-Joining-Algorithmus' generiert. Mittlere SOM Portraits sind für jede taxonomische Gruppe aufgezeigt. Dabei beziehen sich rote Regionen auf Bereiche im Spektrum mit hohen Intensitäten. (b) Dendrogramm der Drosophila-Taxonomie (aus [FGK+10], bearbeitet).

5 Quintessenz

Methodische Aspekte der SOM Portraitierung wurden hier kurz vorgestellt. Sie zielen auf die gezielte Extraktion von Informationen aus großen und komplexen molekularbiologischen Datensätzen ab und stellen die Datenlandschaft sowie die Analyseergebnisse in ausgeklügelter Art und Weise dar. Die individuellen Portraits der Proben sind sehr elegant, da sie unverkennbare Fingerabdrücke der darunterliegenden Datenlandschaft darstellen. Dies ermöglicht eine dem Menschen intuitive Bild-basierte Sichtweise und fördert so die datengetriebene Analyse der Daten und die Generierung von Hypothesen. Die Dimension der ursprünglichen Daten wird stark reduziert und komplexe Meta-Daten für nachfolgende Analysen generiert. Dabei wird die Informationsvielfalt bewahrt und so exploratives und multivariates Vergleichen der Proben ermöglicht.

Literatur

[FGK+10] Ralph Feltens, Renate Görner, Stefan Kalkhof, Helke Gröger-Arndt und Martin von Bergen. Discrimination of different species from the genus Drosophila by intact protein profiling using matrix-assisted laser desorption ionization mass spectrometry. *BMC evolutionary biology*, 10:95, Januar 2010.

[LAT+08] Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, Luigi L Cavalli-Sforza und Richard M Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science (New York, N.Y.)*, 319(5866):1100–4, Februar 2008.

[LK06] Caroline M Li und Robert R Klevecz. A rapid genome-scale response of the transcriptional oscillator to perturbation reveals a period-doubling path to phenotypic change. *Proceedings of the National Academy of Sciences of the United States of America*, 103(44):16254–9, Oktober 2006.

[RGK+10] David Reich, Richard E Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y Durand, Bence Viola, Adrian W Briggs, Udo Stenzel, Philip L F Johnson, Tomislav Maricic, Jeffrey M Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V Shunkov, Anatoli P Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin und Svante Pääbo. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–60, Dezember 2010.

[TKRM05] Benjamin P Tu, Andrzej Kudlicki, Maga Rowicka und Steven L McKnight. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science (New York, N.Y.)*, 310(5751):1152–8, November 2005.

[WLvBB11] Henry Wirth, Markus Loffler, Martin von Bergen und Hans Binder. Expression cartography of human tissues using self organizing maps. *BMC Bioinformatics*, 12(1):306, 2011.

[WvBB12] Henry Wirth, Martin von Bergen und Hans Binder. Mining SOM expression portraits: feature selection and integrating concepts of molecular function. *BioData mining*, 5(1):18, Oktober 2012.

Dr. Henry Wirth, geboren am 05.10.1982 in Meerane, studierte von 2002 bis 2008 Informatik mit Vertiefungsrichtung 'Künstliche Intelligenz' an der TU Chemnitz. Nach bestandenem Diplom erhielt er ein Doktorats-Stipendium vom Helmholtz-Zentrum für Umweltforschung in Leipzig in Kooperation mit dem Interdisziplinären Zentrum für Bioinformatik der Universität Leipzig. Seine Promotion wurde 2012 mit dem Prädikat 'summa cum laude' vollzogen. Seitdem ist er als Postdoc am Zentrum für Bioinformatik tätig und ist dabei unter anderem in verschiedene Projekte im Kontext von Gliomen und Lymphomen involviert.