ETH zürich Research Collection

Doctoral Thesis

Biomedical informatics technologies enabling comprehensive surfaceome analysis

Author(s):
Omasits, Ulrich

Publication Date:
2015

Permanent Link:

https://doi.org/10.3929/ethz-a-010542074 →

Rights / License:

In Copyright - Non-Commercial Use Permitted →

This page was generated automatically upon download from the ETH Zurich Research Collection. For more information please consult the Terms of use.

ETH Library

DISS. ETH NO. 22712

BIOMEDICAL INFORMATICS TECHNOLOGIES ENABLING COMPREHENSIVE SURFACEOME ANALYSIS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

ULRICH OMASITS

MSc, University of Vienna, Austria

BSc, Vienna University of Technology, Austria

born on 19.02.1985

citizen of Austria

accepted on the recommendation of

Prof. Dr. Ruedi Aebersold, examiner

Prof. Dr. Bernd Wollscheid, co-examiner

Prof. Dr. Manfred Claassen, co-examiner

Dr. Christian H Ahrens, co-examiner

2015

Summary

The technology-driven ability to generate increasingly comprehensive biomedical data sets at the genome, transcriptome and proteome level has shifted the bottleneck to their subsequent integration, analysis and mining – posing a major challenge for the research community at large. This thesis addresses the biomedical informatics challenge and provides solutions for integrative data analysis and visualization with a particular focus on the cell surface proteome. The repertoire of cell surface exposed proteins, referred to as surfaceome, acts as a cellular signaling gateway connecting the extracellular signaling space to intracellular signaling networks. As such it enables, but also limits cellular interactions with the microenvironment. Due to the functional importance of the surfaceome it is of high interest for basic research and the biopharmaceutical industry. In fact, the majority of currently available drugs target cell surface proteins. However, despite its functional importance and extracellular accessibility the surfaceome surprisingly remains largely uncharted *terra incognita*. It is currently not known which proteins of the proteome can actually reside in the plasma membrane at a particular point in time and how they are distributed over this landscape. This gap of knowledge is mainly due to technical limitations in surveying the surfaceome comprehensively in space and time. Recent advances in chemoproteomic technologies in combination with biomedical informatics technologies enable now to bridge this gap and provide an opportunity to establish a mechanistic understanding of the surfaceome as an integrated signaling gateway of the cell responding to complex environmental cues.

In this thesis, I utilized these recent bioanalytical advances in combination with my biomedical informatics expertise to develop and apply reusable bioinformatic frameworks which enabled for the first time the description of complete surfaceomes, representing a critical need for the ability to study and target entire surfaceomes with quantitative approaches.

In chapters two and three, a data-driven experimental strategy for identification of the complete proteome of a prokaryotic organism was developed and applied to *Bartonella henselae*. The combined approach of experimental and machine learning methods enabled inference of protein subcellular location and definition of the organism's surfaceome. In two conditions that mimic interaction with the prokaryote's major host organisms, the surfaceome was shown to be most strongly regulated.

In chapters four and five, surfaceome data sets of 41 human and 31 mouse cell types were experimentally established by a team effort to generate the most comprehensive experimental resource of the human and mouse surfaceome to date. Based on this unique experimentally verified surfaceome resource, a biomedical informatics approach was developed, building a computational model that was able to learn the characteristic properties of human cell surface proteins. This computational model was then applied to predict potential surface proteins from the human genome that have not been associated with the cell surface location and function before. The combination of high-quality experimental surfaceome data sets with advanced biomedical informatics methods led to the first comprehensive description of the human surfaceome consisting of 2886 cell surface

proteins, i.e. 14 % of the human proteome. With two out of three of currently available drugs targeting only 408 of these cell surface proteins, this particular subset of the human proteome holds enormous potential for development of novel diagnostic and therapeutic drug targets and for biomarkers of disease.

In chapter six to eight, three broadly applicable software tools were developed in order to efficiently analyze, process and visualize the massive data sets generated in the course of the presented surfaceome studies. All tools developed were made publicly accessible via web interfaces that are already widely used by the research community. PeptideRank (http://wlab.ethz.ch/peptiderank) is used to predict the best-observable proteotypic peptides for any given protein in a mass spectrometry-based workflow. Protter (http://wlab.ethz.ch/protter) is an open-source tool for visualization of individual cell surface proteoforms in the context of current knowledge, annotations, experimental evidence and predictions. Meteor (http://wlab.ethz.ch/meteor) is an automated, high-throughput mass spectrometry-based proteomics analysis pipeline.

In conclusion, a comprehensive characterization and definition of the surfaceome was established, from simple prokaryotes to complex eukaryotic cells. An innovative approach based on biomedical informatics methods was developed to map out and predict this hitherto *terra incognita* on cellular surfaces. The presented functional analyses and surfaceome resources harbor an enormous potential in the biopharmaceutical industry for discovery of novel drug targets, biomarkers of disease, and therapeutics for infectious diseases. In basic research, the strategies and tools developed will provide a guiding role for studies assessing the plasticity and response of cell surface proteomes in response to differentiation or perturbations, aiming at a mechanistic understanding of how cells sense and communicate with their microenvironment.

Zusammenfassung

Der technologische Fortschritt ermöglicht es immer umfassendere biomedizinische Genom, Transkriptom und Proteom Daten zu generieren, wodurch der limitierende Faktor in der Forschung auf die Analyse, Integration und Auswertung der Daten verschoben wurde – eine große Herausforderung für die wissenschaftliche Gemeinschaft. Meine Doktorarbeit widmet sich dieser Herausforderung der biomedizinischen Informatik und erarbeitet Methoden für die integrative Datenanalyse und -visualisierung mit speziellem Fokus auf das Zelloberflächenproteom. Das Repertoire von an der Zelloberfläche exponierten Proteinen, auch Surfaceom genannt, agiert für die Zelle als Schnittstelle zwischen dem extrazellulären Raum und dem intrazellulären Signalnetzwerk. Dadurch ermöglicht es und limitiert gleichzeitig die Interaktion der Zelle mit ihrer Mikroumgebung. Aufgrund dieser wichtigen Funktion ist das Surfaceom von größtem Interesse für die Grundlagenforschung aber auch für die Pharmaindustrie. Tatsächlich zielt bereits eine Mehrheit der heutigen Medikamente auf Zelloberflächenproteine ab. Trotz der Relevanz und auch der Zugänglichkeit bleibt ein erstaunlich großer Teil des Surfaceoms weitestgehend unkartierte *Terra incognita*. Es ist ungewiss welche Proteine sich zu einem gewissen Zeitpunkt in der Plasmamembran befinden und wie sie dort verteilt sind. Diese Wissenslücke resultiert vor allem von technischen Einschränkungen das Surfaceom umfassend in Raum und Zeit zu untersuchen. Jüngste Entwicklungen in chemo-proteomischen Technologien und Methoden der biomedizinischen Informatik ermöglichen es, diese Einschränkungen zu überwinden und erlauben ein mechanistisches Verständnis des Surfaceoms, als eine integrierte Signalisationsschnittstelle die auf komplexe Signale der Umgebung reagiert.

In der vorliegenden Doktorarbeit habe ich diese jüngsten Fortschritte der Analysemethoden zusammen mit meiner Expertise der biomedizinischen Informatik eingesetzt um Ansätze zu entwickeln mit denen zum ersten Mal vollständige Surfaceome beschrieben werden können – eine Voraussetzung für die gezielte, quantitative Analyse von Surfaceomen.

In Kapitel zwei und drei wurde eine experimentelle Strategie zur Identifizierung des kompletten Proteoms eines Prokaryoten entwickelt, am Beispiel von *Bartonella henselae*. Eine Kombination von experimentellen Methoden und Methoden des maschinellen Lernens ermöglichte die Bestimmung der subzellulären Lokalisierung von allen Proteinen und damit auch die Bestimmung des Surfaceoms. In zwei Konditionen, die die Interaktion des Prokaryoten mit seinen Wirten nachstellen, konnte gezeigt werden, dass Zelloberflächenproteine am stärksten reguliert waren.

In Kapitel vier und fünf wurden experimentelle Surfaceom Daten von 41 humanen und 31 Maus Zelltypen generiert, in einem gemeinsamen Bestreben nach der umfassensten experimentellen Ressource des Human- und Maus-Surfaceoms. Basierend auf dieser einzigartigen, experimentell validierten Surfaceom Ressource, wurde ein Ansatz der biomedizinischen Informatik entwickelt der mittels eines Computermodells die charakteristischen Eigenschaften von Zelloberflächenproteinen erlernen konnte. Mit diesem Modell wurden in Folge, basierend auf dem humanen Genom,

potenzielle Zelloberflächenproteine vorhergesagt die bisher noch nicht mit der Zelloberfläche assoziiert wurden. Die Kombination von qualitativ hochwertigen experimentellen Surfacom Daten und fortschrittlichen Methoden der biomedizinischen Informatik führten zur ersten umfassenden Beschreibung des humanen Surfacoms, welches aus 2886 Zelloberflächenproteinen besteht, was 14 % des Proteoms entspricht. Da zwei von drei der heutigen Medikamente auf lediglich 408 dieser Zelloberflächenproteine abzielen, ist dieser spezielle Teil des humanen Proteoms äußerst vielversprechend für die Entwicklung neuer diagnostischer und therapeutischer Medikamente und Biomarker.

In den Kapiteln sechs bis acht wurden drei breit einsetzbare Software-Werkzeuge entwickelt um die massenhaften Daten der Surfacom Studien effizient analysieren, prozessieren und visualisieren zu können. Jede Software wurde auf einer Internetseite öffentlich zugänglich gemacht und wird von der wissenschaftlichen Gemeinschaft rege benutzt. PeptideRank (http://wlab.ethz.ch/peptiderank) macht Vorhersagen zu den massenspektrometrisch am besten beobachtbaren proteotypischen Peptiden eines Proteins. Protter (http://wlab.ethz.ch/protter) ist eine Open-Source Software zur Visualisierung von individuellen Zelloberflächenproteinen, im Kontext von Annotationen, experimentellen Daten und Vorhersagen. Meteor (http://wlab.ethz.ch/meteor) ist ein Hochdurchsatz-Analysewerkzeug für Daten der Massenspektrometrie-basierten Proteomik.

Insgesamt wurde das Surfaceom umfassend charakterisiert und definiert, für einfache Prokaryoten wie auch für komplexere eukaryotische Zellen. Ein innovativer Ansatz wurde entwickelt, basierend auf Methoden der biomedizinischen Informatik, um die bisherige *Terra incognita* des Surfaceoms zu kartieren. Die durchgeführten funktionalen Analysen und die Surfaceom Ressource beinhalten ein enormes Potential für die Pharmaindustrie um neue Ansatzpunkte für Medikamente zu entwickeln und Biomarker für Krankheiten zu finden. In der Grundlagenforschung werden die entwickelten Strategien und Werkzeuge zukünftige Surfaceom Studien maßgeblich beeinflussen und es ermöglichen, das Verhalten und die Plastizität des Zelloberflächenproteoms in Reaktion auf Differenzierung oder Perturbation zu untersuchen, mit dem Ziel eines mechanistischen Verständnisses wie Zellen kommunizieren und ihre Mikroumgebung wahrnehmen.