Proteomik

DOI: 10.1002/ange.201406545

Vom humanen Genom zum humanen Proteom**

Javier Muñoz und Albert J. R. Heck**

Humanes Genom · Humanes Proteom · Massenspektrometrie · Proteomik

Eines der grundlegenden Ziele der Biologie ist es, vollständig zu verstehen, wie eine Zelle oder sogar ein kompletter Organismus funktioniert. Idealerweise könnte solches Wissen genutzt werden, um die zelluläre Antwort auf spezifische Signale oder Krankheiten vorherzusagen. Ein erster Schritt hin zu diesem Ziel ist die Identifizierung und Charakterisierung aller in den Zellen vorhandenen molekularen Akteure. Das Humangenomprojekt[1] war vermutlich eines der bisher ambitioniertesten wissenschaftlichen Unterfangen und löste die ersten Teile dieses Puzzles. Die Aufklärung der drei Milliarden Basenpaare, die unsere DNA bilden, wurde weltweit begeistert aufgenommen, da diese Informationen zum Verständnis der molekularen Mechanismen menschlicher Krankheitsbilder führen könnten.

Bald darauf wurde man sich jedoch der unserem genetischen Kode innewohnenden Komplexität bewusst. Eine überraschende Entdeckung war der geringe Anteil proteinkodierender DNA (weniger als 2% des Genoms) mit ca. 20000 humanen Genen. Gleichwohl lassen aktuelle Analysen darauf schließen, dass 80% des menschlichen Genoms funktionell sind, also entweder transkribiert werden, regulatorische Proteine binden oder mit anderen biochemischen Funktionen in Zusammenhang stehen.[2] Zwar sind Informationen zum Genom unerlässlich, jedoch werden hierbei die wichtigsten molekularen Effektoren in Zellen – die Proteine – außer Acht gelassen. Jeder Wissenschaftler wird zustimmen, dass die Analyse des Proteoms relevanter ist, aber wegen technischer Hindernisse und der um mehrere Größenordnungen höheren Komplexität des Proteoms bisher in geringerem Maße durchgeführt wurde. Während das Genom jeder Zelle des menschlichen Körpers nahezu identisch und im Lauf des Lebens eines Organismus relativ konstant ist, sind die Proteome der Zellen sehr unterschiedlich und verändern sich beträchtlich im Verlauf der Zeit (Abbildung 1).

Ungeachtet dieser Schwierigkeiten hat sich das Gebiet der Proteomforschung im letzten Jahrzehnt enorm weiterentwickelt, vor allem dank Fortschritten bei Massenspektrometrie und Bioinformatik, und schließt nun in gewisser Weise zur Genom- und Transkriptomforschung auf. Dies belegen zwei aktuelle Veröffentlichungen in *Nature* von einer deutschen Gruppe unter der Leitung Bernhard Küsters[3] und einem von Akhilesh Pandey koordinierten, US-amerikanisch/indischen Gemeinschaftsprojekt,[4] die unabhängig voneinander beispiellose Versuche unternommen haben, alle im Genom kodierten menschlichen Proteine nachzuweisen. Hierfür haben beide Labore umfangreiche Proteomanalysen von mehr als 70 menschlichen Geweben und Körperflüssigkeiten sowie mehr als 150 Zelllinien durchgeführt. Zwar haben beide Gruppen sehr ähnliche, Massenspektrometrie-basierte Arbeitsabläufe genutzt, es gibt allerdings einige Unterschiede zwischen ihren Studien, besonders hinsichtlich der Analysentiefe. Während Pandey et al. ungefähr 2000 massenspektrometrische (LC-MS-)Analysen durchgeführt haben, machten Küster et al. mehr als 6000 Analysen und nutzten die Daten von 10000 weiteren Messungen aus öffentlich zugänglichen Proteomikarchiven. Unter Annahme einer durchschnittlichen Laufzeit von 2 h würde die zur Aufzeichnung dieser Daten benötigte Messzeit beeindruckende 34000 h betragen (4.3 Jahre, wenn nur ein Massenspektrometer genutzt würde). Die Auswertung aller Daten führte zur Identifizierung von 946000 bzw. 293000 nichtredundanten einzigartigen Peptidsequenzen in Küsters bzw. Pandeys Studien. Erstaunlicherweise (und trotz des signifikanten Tiefenunterschiedes) wurde in beiden Studien eine nahezu identische Zahl proteinkodierender Gene nachgewiesen: 18097 (Küster) und 17294 (Pandey). Auch wenn ein sorgfältiger Vergleich beider Arbeiten noch aussteht, lässt sich bereits eines schlussfolgern: das unzweifelhafte Vorhandensein von Proteintranslation für 90–95% der menschlichen Gene. Dies ist ein außerordentlich bedeutender Befund, da fast ein Drittel der humanen Gene bislang kaum annotiert war und kein experimenteller Nachweis bestand, dass sie Proteine ergeben könnten. Eine weitere relevante Erkenntnis aus diesen Studien betrifft das Ausmaß alternativen Spleißens bei der Generierung von Proteinisoformen. Es ist offen-

[*] Prof. Dr. A. J. R. Heck
Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University
Padualaan 8. 3584 CH Utrecht (Niederlande)
E-Mail: a.j.r.heck@uu.nl

Dr. J. Muñoz
Proteomics Unit
Spanish National Cancer Research Centre (CNIO, ProteoRed-ISCIII)
Melchor Férnandez Almagro, 3, 28029 Madrid (Spanien)
E-Mail: jmunozpe@cnio.es

[**] A.J.R.H. dankt für Unterstützung durch das Netherlands Proteomics Center, das von der Niederländischen Organisation für Wissenschaftliche Forschung (NWO) geförderte Proteomik-Großprojekt Proteins@Work (Projekt 184.032.201) sowie das PRIME-XS-Projekt (Förderungsnummer 262067), das durch das 7. Rahmenprogramm der Europäischen Union finanziert wird. J.M. wird durch das Ramón y Cajal Programm (MINECO) RYC-2012-10651 unterstützt. Die CNIO-Proteomikgruppe gehört zu ProteoRed, PRB2-ISCIII, Förderungsnummer PT13/0001.

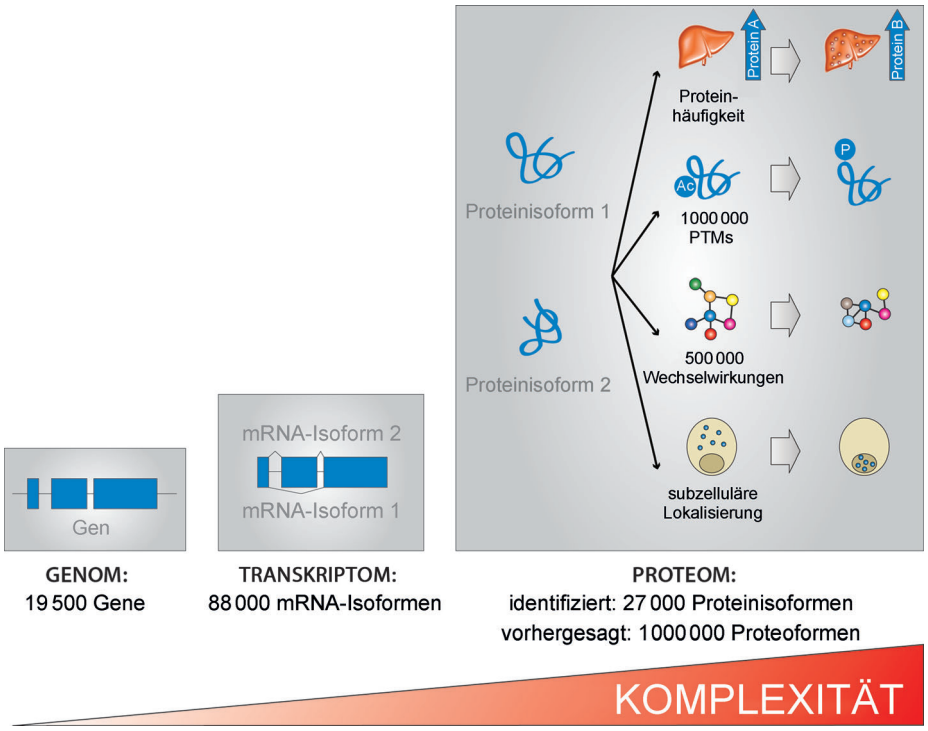

Abbildung 1. Ein Gen, ein Protein? Die Komplexität des humanen Proteoms: Die Zahl humaner proteinkodierender Gene wird auf ca. 20 000 geschätzt. Aufgrund alternativen Spleißens ist bereits die Zahl transkribierter mRNAs signifikant größer. Die in zwei aktuellen Studien veröffentlichten ersten Entwürfe des humanen Proteoms belegen die Translation von 18 000 Proteinen (mit mehr als 27 000 Isoformen). Gewebsspezifische Expression, posttranslationale Modifizierungen (PTMs), Protein-Protein-Wechselwirkungen und spezifische subzelluläre Lokalisierung tragen zur zusätzlichen Steigerung der Proteomkomplexität bei. Diese Proteomprofile werden infolge biologischer oder pathologischer Störungen dynamisch modifiziert. Schätzungen gehen von mehr als 1 000 000 Proteoformen (resultierend aus genetischer Variation, alternativem Spleißen und PTMs) im menschlichen Körper aus.

sichtlich, dass die Zahl der Gene nicht mit der Komplexität eines Organismus korreliert (*C. elegans* hat z. B. 20 500 Gene) und es wurde vermutet, dass alternatives Spleißen das Repertoire funktioneller Proteine vergrößern könnte. In den Proteomikstudien konnten jedoch nur 9000 der 67 000 in Uniprot annotierten Isoformen identifiziert werden. Zwar erzeugen einige dieser Isoformen nur ein einzigartiges Peptid, was die Wahrscheinlichkeit der Identifizierung durch Proteomanalytik verringert, dennoch könnten diese Daten die Vorstellung untermauern, dass es eine dominante Isoform pro Gen gibt.[5] Beide Studien bestätigen die Existenz eines in allen Geweben vorhandenen Kernproteoms, das aus „Haushaltsproteinen“ besteht (z. B. Histone, ribosomale und zytoskelettale Proteine), die mehr als 75 % der Gesamtproteinmasse ausmachen. Andererseits wurde die Expression vieler gewebsspezifischer Proteine beobachtet, die Gewebsproteomsignaturen definieren. Alle diese Daten wurden in zwei nutzerfreundlichen Portalen (https://www.proteomicsdb.org und www.humanproteomemap.org) öffentlich zugänglich gemacht. Somit ist es Wissenschaftlern möglich, diese ausgedehnten Entwürfe humaner Proteome zu durchsuchen.

Durch die Entwicklung von Techniken zur Hochdurchsatzsequenzierung von DNA und RNA wurde die Analyse von Genomen und Transkriptomen mit akzeptablem Zeit- und Kostenaufwand ermöglicht. Dies ist zum Teil in der Beschaffenheit von DNA und RNA begründet, die aus nur vier durch Polymerasen einfach zu vervielfältigenden Nukleotiden aufgebaut sind. Dagegen sind Proteome aus analytischer Sicht deutlich anspruchsvoller. Beispielsweise haben die 20 Aminosäuren, die als Proteinbausteine fungieren, recht unterschiedliche physikochemische Eigenschaften. Darüber hinaus kann die Proteinhäufigkeit in Zellen um zehn Größenordnungen variieren, und Protein-Protein-Wechselwirkungen sowie posttranslationale Modifizierungen (PTMs), z. B. Phosphorylierung oder Glycosylierung, ändern sich hochdynamisch in Raum und Zeit. Gleichwohl entwickelt sich, parallel zur Gen- und Transkriptsequenzierung, eine Proteomik „der nächsten Generation“, die vorwiegend auf bedeutenden Verbesserungen im Bereich der analytischen Chemie (Probenvorbereitung und -auftrennung), Informatik, Massenspektrometrie und Datenauswertung beruht und die Analyse komplexer Proteome zu relativ niedrigen Kosten sowie mit hoher Geschwindigkeit ermöglicht.[6] Der üblichste Arbeitsablauf zur Proteomanalyse besteht aus 1) Proteinextraktion, 2) proteolytischer Spaltung, 3) Peptidauftrennung, 4) massenspektrometrischer Messung und 5) Datenauswertung (Abbildung 2). Eine effiziente Peptidauftrennung ist entscheidend, um eine tiefe Proteomabdeckung zu erlangen, da 1–2 Millionen Peptide infolge der proteolytischen Spaltung eines menschlichen Proteoms generiert werden. Mehrdimensionale Strategien, die orthogonale Auftrennungstechniken miteinander kombinieren, sind essenziell. Um die Peptide bestmöglich aufzutrennen, arbeiten HPLC-Systeme mittlerweile bei sehr hohem Druck (15 000 psi) und

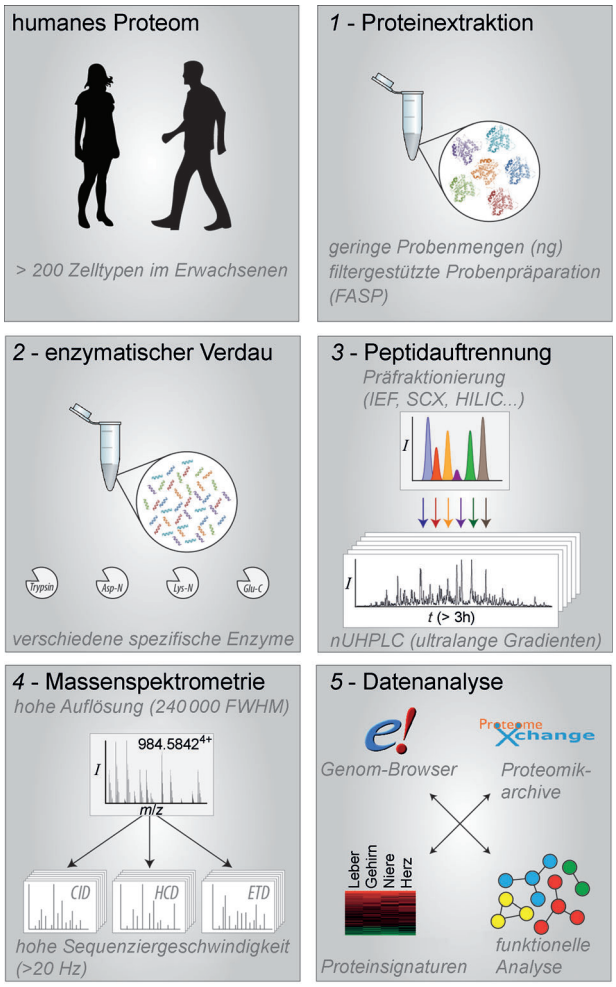

Abbildung 2. Typischer Arbeitsablauf zur Analyse hochkomplexer humaner Proteome. Einige aktuelle technische Entwicklungen sind in Grau aufgeführt. CID = stoßinduzierte Dissoziation, ETD = Elektronentransfer-Dissoziation, HCD = stoßinduzierte Dissoziation bei erhöhter Energie, IEF = isoelektrische Fokussierung, HILIC = hydrophile Interaktionsflüssigchromatographie, nUHPLC = Nano-Ultrahochleistungs-Flüssigkeitschromatographie, SCX = Ionenaustauschchromatografie mit starkem Kationenaustauscher.

nutzen lange Säulen (> 50 cm) mit kleinem innerem Durchmesser (< 50 µm), die mit kleinen Partikeln (1.7 µm) gepackt sind. Dadurch wird die Probenkomplexität bedeutend reduziert, den dynamischen Bereich betreffende Probleme werden folglich verringert, und die nachfolgende Peptidsequenzierung mit Massenspektrometrie wird vereinfacht. Massenspektrometer haben sich beträchtlich entwickelt: Moderne Instrumente verfügen über hohe Auflösungsvermögen von 240 000 FWHM (FWHM = Halbwertsbreite), um Vorläufer- und Fragment-Ionen mit hoher Massengenauigkeit zu messen, sowie über höhere MS/MS-Peptidsequenzierungsgeschwindigkeiten von 20 Hz. Verschiedene Fragmentierungstechniken (z. B. CID, HCD und ETD) können zur spezifischen Peptidsequenzierung verwendet werden, wodurch sich die Identifizierungsquote erhöht. Nicht außer Acht gelassen werden sollte die parallele Entwicklung im Bereich der Informatik, mit einer Fülle erhältlicher Softwarepakete, die Gigabytes von Roh-MS-Daten in Tausende identifizierte Peptide umwandeln können. Das „Hefeproteom-Wettrennen“ veranschaulicht diesen Fortschritt: Zur Identifizierung der 4000 Hefeproteine wurde 2008 eine 144-stündige MS-Analyse benötigt,[7] die 2013 auf eine Stunde verkürzt werden konnte.[8] Zusammen haben diese parallelen Fortschritte in den grundlegenden Techniken nun jene ersten Skizzen des humanen Proteoms möglich gemacht.

Die vorgestellten Karten bieten umfangreiche Bezugsquellen für das humane Proteom. Dennoch ist es noch ein langer Weg bis zum vollständigen Verständnis des Proteoms. Die Erkenntnis, dass fast jedes Gen auf Proteinebene manifestiert ist, verrät uns noch nichts über die Funktion dieser Proteine und deren Regulation durch PTMs und dynamische Wechselwirkungen, die sie mit anderen molekularen Entitäten in der Zelle eingehen. Ein logischer nächster Schritt ist die Identifizierung aller PTMs (mehr als 200 Arten sind bekannt[9]), die insbesondere menschlichen Proteinen anhaften, um zu verstehen, wie sie deren Funktionen, Aktivitäten und/oder Lokalisierungen regulieren. Gleichermaßen würde die gewebsspezifische Rekonstruktion aller Protein-Protein-Wechselwirkungen Proteinnetzwerke erzeugen, die zur Identifizierung neuer funktioneller Komplexe und Signalwege beitragen würden. Alternative Proteomikstrategien werden in diesem Bereich nötig sein. In Top-down-Ansätzen werden intakte Protein durch MS ohne vorherige Proteolyse analysiert, sodass genauere Kenntnis über Proteoformen und die gegenseitige Abhängigkeit von PTMs erlangt wird.[10] Des Weiteren kann Antikörper-basiertes Profiling zur Entschlüsselung lokaler Proteinexpression in verschiedenen Zelltypen genutzt werden.[11] Schließlich wird die Analyse der dynamischen Umbildung des Proteoms infolge biologischer Störungen essenzielle Informationen zum Verständnis der Biologie menschlicher Zellen liefern. Die Verfügbarkeit hoch relevanter, großer Proteomikdatensätze wird, gemeinsam mit anderen „-omik“-Daten, der Schlüssel zum Erfolg personalisierter Medizin (Präzisionsmedizin) sein.

Eingegangen am 24. Juni 2014
Online veröffentlicht am 30. Juli 2014

[1] a) E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. Fitz Hugh et al., *Nature* **2001**, *409*, 860 – 921; b) J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt et al., *Science* **2001**, *291*, 1304 – 1351.
[2] E. Pennisi, *Science* **2012**, *337*, 1159 – 1161.
[3] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, et al., *Nature* **2014**, *509*, 582 – 587.
[4] M.-S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain et al., *Nature* **2014**, *509*, 575 – 581.
[5] M. Gonzàlez-Porta, A. Frankish, J. Rung, J. Harrow, A. Brazma, *Genome Biol.* **2013**, *14*, R70.
[6] a) A. F. M. Altelaar, J. Munoz, A. J. R. Heck, *Nat. Rev. Genet.* **2013**, *14*, 35 – 48; b) A. Bensimon, A. J. Heck, R. Aebersold, *Annu. Rev. Biochem.* **2012**, *81*, 379 – 405; c) J. Cox, M. Mann, *Annu. Rev. Biochem.* **2011**, *80*, 273 – 299; d) J. R. Yates, C. I.

Ruse, A. Nakorchevsky, *Annu. Rev. Biomed. Eng.* **2009**, *11*, 49 – 79.

[7] L. M. F. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Fröhlich, T. C. Walther, M. Mann, *Nature* **2008**, *455*, 1251 – 1254.

[8] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, J. J. Coon, *Mol. Cell. Proteomics* **2014**, *13*, 339 – 347.

[9] C. T. Walsh, S. Garneau-Tsodikova, G. J. Gatto, *Angew. Chem.* **2005**, *117*, 7508 – 7539; *Angew. Chem. Int. Ed.* **2005**, *44*, 7342 – 7372.

[10] J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman, K. R. Durbin, J. D. Tipton, A. Vellaichamy, J. F. Kellie, M. Li et al., *Nature* **2011**, *480*, 254 – 258.

[11] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober et al., *Nat. Biotechnol.* **2010**, *28*, 1248 – 1250.