Maschinelle Lernmethoden zur Analyse von Tiling-Array-Daten

Georg Zeller

Europäisches Laboratorium für Molekularbiologie
Meyerhofstr. 1, 69117 Heidelberg, Deutschland
georg.zeller@gmail.com

Abstract: Im Rahmen meiner Dissertation [Zel10] entwickelte ich auf Maschinellen Lerntechniken basierende, bioinformatische Methoden, um zur Beantwortung zentraler molekularbiologischer Fragestellungen beizutragen:

- In welchen Bereichen des Genoms unterscheiden sich einzelne Individuen derselben Spezies?
- Welche Bereiche des Genoms beinhalten Gene, und in welchen Zellen, Organen und Entwicklungsstadien werden diese in mRNA-Moleküle transkribiert?

Diese beiden Probleme weisen einige – vielleicht unerwartete – Gemeinsamkeiten auf: Erstens lassen sich beide als Segmentierungsprobleme formalisieren. Zweitens hat die Molekularbiologie eine sehr flexible Hochdurchsatz-Experimentiertechnik entwickelt, sogenannte Tiling-Arrays (bzw. deren Weiterentwicklung zu Resequencing-Arrays), die es ermöglich, diese beiden (und weitere) Fragestellungen experimentell zu bearbeiten. Im wesentlichen liefert diese Technik eine Sequenz von Messwerten, die in einem regelmäßigen Raster das gesamte Genom abdecken. Das Segmentierungsproblem bei der Analyse dieser Sequenzdaten besteht nun darin, die Teilbereiche zu erkennen, welche dem gesuchten biologischen Phänomen entsprechen, nämlich einerseits variable Genomregionen (im Unterschied zu solchen, wo sich Individuen nicht unterscheiden) und andererseits Segmente, aus denen mRNA-Moleküle generiert werden.

Zur Lösung dieser Probleme entwickelte ich Segmentierungsmethoden, die auf der sogenannten Hidden Markov Support Vector Machine (HMSVM) basieren und sich durch folgende Eigenschaften auszeichnen:

- *Genauigkeit* der Vorhersagen war von entscheidender Bedeutung, da meine Resultate die Grundlage für weitergehende experimentelle Forschung bildeten. Wo vergleichbare Methoden verfügbar waren, konnte ich die stark verbesserte Genauigkeit der neu entwickelten Lernmethoden belegen.
- Ich untersuchte empirisch, dass die hohe Genauigkeit teils einem ausgefeilten Modellierungsansatz und teils einem neuen *diskriminativen Lernalgorithmus* mit großer *Robustheit gegen Rauschen* zugeschrieben werden kann. Angesichts des starken Rauschens in Tiling-Array-Daten erwies sich Robustheit als Schlüsseleigenschaft.
- Ein weiterer Schwerpunkt lag auf der *Effizienz* der Methoden. Analysen ganzer Genome erfordern schnelle Vorhersagealgorithmen, und angesichts langer Trainingssequenzen sind Lernmethoden im Vorteil, die bereits anhand weniger Trainingsbeispiele in der Lage sind, genaue Vorhersagen zu machen.
- Die Verwandschaft zu Hidden Markov Modellen (HMMs) mit einem *breiten Anwendungsspektrum* in der Bioinformatik eröffnet für die Anwendung der HMSVM viele Möglichkeiten über die hier beschriebenen hinaus.

1 Maschinelles Lernen in den Lebenswissenschaften

Hochdurchsatzverfahren in der Molekularbiologie, wie z.B. die DNA-Sequenzierung, haben in den letzten Jahren entscheidend unsere Sicht auf uns selbst und die Funktionsweise unseres Körpers verändert [LLB+01]. Die automatisierte Datengenerierung im großen Stil, die heute fast alle molekularbiolgischen Forschungsfelder prägt, sorgt für einen stetig wachsenden Bedarf an „intelligenten“ Analysetechniken, um aus riesigen, oft komplexen und heterogenen Datenmengen biologische Einsichten zu gewinnen. Die Entwicklung solcher bioinformatischer Methoden, die Konzepte aus den Bereichen des Maschinellen Lernens und des verwandten Data Minings entlehnen und dabei besonders dem biologischen Kontext Rechnung tragen, bildet ein faszinierendes, interdisziplinäres Forschungsgebiet.

Maschinelle Lernverfahren „lernen“ eine bestimmte Eigenschaft von Beispieldaten (d.h. sie erkennen verallgemeinerbare Gesetzmäßigkeiten), um diese dann auf neuen, unbekannten Daten vorherzusagen. Handelt es sich um die Vorhersage einer Klassenzugehörigkeit, spricht man von Klassifikation. Wenn jedoch jedes Lernbeispiel selbst aus einer Sequenz von Messpunkten (oder allgemein aus einer Merkmalssequenz) besteht, ist man oft an einer Segmentierung dieser Sequenz in bestimmte Teilbereiche interessiert.

In der Bioinformatik sind wir häufig mit *Segmentierungsproblemen* konfrontiert. Genvorhersage beispielsweise, d.h. die Segmentierung des Genoms in Gene, in welchen sich wiederum exonische Bereiche mit intronischen Bereichen abwechseln, sowie in umgebende intergenische Bereiche, ist ein bioinformatisches Kernproblem (Abb. 1) [SZZ+09]. Dieses und andere Segmentierungsprobleme lassen sich darauf reduzieren, jeder Position in einer langen Sequenz ein atomares Label zuzuweisen (z.B. exonisch, intronisch oder intergenisch). Oft existieren hier Abhängigkeiten (z.B. sind Introns im Genom stets von Exons umgeben), die es nicht erlauben, einfach Klassifikationsverfahren positionsweise unabhängig anzuwenden. Derartige Labelabhängigkeiten lassen sich jedoch mithilfe eines Zustandsübergangsmodells formal beschreiben (Abb. 1).

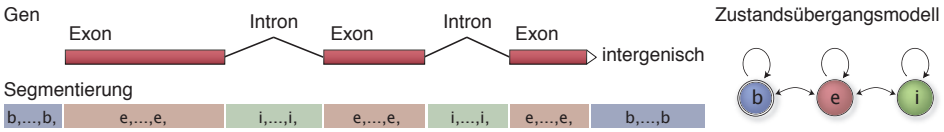

Abbildung 1: (links) Segmentierung eines Teilbereichs des Genoms in intergenische (b), exonische (e) und intronische (i) Bereiche (farbkodiert) entsprechend dem schematisch dargestellten Gen. (rechts) Labelabhängigkeiten, wie die, dass Intron-Segmente (grün, i) stets von Exon-Segmenten (rot, e) umgeben sein müssen, lassen sich mithilfe eines Zustandsübergangsmodells formalisieren.

Ein wichtiger Aspekt vieler Bioinformatikprobleme ist, dass wir an Ergebnissen für große Datensätze, meist für ein ganzes Genom, interessiert sind. Bei der Entwicklung Maschineller Lernverfahren zur Genomanalyse – für höhere Organismen in der Größe von >100 Mio. Informationseinheiten – ist daher *Skalierbarkeit* und Parallelisierbarkeit von großer Wichtigkeit.

2 Diskriminative Segmentierungsverfahren

Die klassische, adaptive und statistisch fundierte Segmentierungsmethode, das sogenannte Hidden Markov Modell (HMM), erfreut sich in der Bioinformatik großer Beliebtheit [DEKM98]. Das HMM ist ein probabilistisches Modell, welches typischerweise mit dem Maximum-Likelihood-Ansatz generativ trainiert wird. Ein neuerer Lernalgorithmus, Hidden Markov Support Vector Machine (HMSVM) genannt [ATH03], basiert stattdessen auf einer *diskriminativen Trainingsmethode*, die wie die SVM [Vap95, SS02] auf dem Maximum-Margin-Prinzip beruht. Die Funktionsweise der HMSVM wird hier kurz skizziert, gefolgt von einer empirischen Evaluation ihrer Eigenschaften im Vergleich zum generativen HMM.

Hidden Markov Support Vector Machines

Formal ist eine Funktion f gesucht, die das Segmentierungsproblem

$$f : X \to \mathcal{S}^\star$$

löst, dadurch dass sie einer Merkmalssequenz $\mathbf{x} \in X$ die richtige Labelsequenz $\boldsymbol{\pi} \in \mathcal{S}^\star$ gleicher Länge zuweist. Hierbei bildet $\mathcal{S}$ die Menge der atomaren Labels. Indem wir eine Diskriminanzfunktion F trainieren, die Güte einer Labelsequenz geeignet zu bewerten, erhalten wir die Funktion f folgendermaßen:

$$f(\mathbf{x}) = \operatorname*{argmax}_{\boldsymbol{\pi} \in \mathcal{S}^\star} F(\mathbf{x}, \boldsymbol{\pi}).$$

Unter der Voraussetzung, dass F die Markoveigenschaft erfüllt, existiert ein auf Dynamischer Programmierung beruhender, effizienter Algorithmus, der unter Berücksichtigung von Labelabhängigkeiten (vgl. Abb. 1) das maximierende $\boldsymbol{\pi}$ berechnet [DEKM98].

Gemäß dem *Maximum-Margin-Prinzip* wird F anhand von n Trainingsbeispielen, für welche die korrekte Segmentierung bekannt ist $(\mathbf{x}^{(i)}, \boldsymbol{\pi}^{(i)}),\ i = 1, \ldots, n$, so optimiert, dass die korrekte Segmentierung $\boldsymbol{\pi}^{(i)}$ am besten bewertet wird, und zwar mit einem *großen Margin* besser als *alle* falschen Segmentierungen $\overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)}$

$$F(\mathbf{x}^{(i)}, \boldsymbol{\pi}^{(i)}) - F(\mathbf{x}^{(i)}, \overline{\boldsymbol{\pi}}) \gg 0 \qquad \forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)} \quad \forall i.$$

Ein wesentlicher Teil meiner Arbeit bestand darin, für die Diskriminanzfunktion F eine geeignete Parametrisierung $\boldsymbol{\theta}$ zu finden. Aus stückweise linearen Funktionen ließ sich ein sehr flexibles Bewertungsmodell konstruieren, welches sich zudem effizient trainieren lässt (Details finden sich in [Zel10, ZCS+08]).

Das führt zum HMSVM-Trainingsproblem (hier in einer vereinfachten Variante):

$$\min_{\boldsymbol{\theta}} \ \Omega(\boldsymbol{\theta})$$

$$\text{s.t.} \qquad F_\theta(\mathbf{x}^{(i)}, \boldsymbol{\pi}^{(i)}) - F_\theta(\mathbf{x}^{(i)}, \overline{\boldsymbol{\pi}}) \ \geq \ \Delta(\boldsymbol{\pi}^{(i)}, \overline{\boldsymbol{\pi}}) \quad \forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)} \quad \forall i = 1, \ldots, n$$

wobei Ω ein Regularisierungsterm ist, der die Modellkomplexität beschreibt. Diese wird hier minimiert um Überanpassung zu vermeiden (Details in [Zel10, ZCS+08])

Da wir sowohl eine lineare Parametrisierung in $F_{\boldsymbol{\theta}}$ als auch einen linearen (oder quadratischen) Regularisierer Ω verwendeten, erhielten wir ein lineares (bzw. quadratisches) Optimierungsproblem. In dieser Formulierung des Problems wird der Margin mit einer Loss-Funktion $\Delta(\boldsymbol{\pi}^{(i)}, \overline{\boldsymbol{\pi}})$ skaliert, was intuitiv bedeutet, dass Vorhersagen, die sich von der richtigen Segmentierung stark unterscheiden, auch weit weniger gut betwertet werden sollen. In meiner Arbeit untersuchte ich im Detail, dass eine sorgfältige Modellierung dieser Loss-Funktion entscheidenden Einfluss auf die Vorhersagequalität des gelernten Modells hat [Zel10].

Die Schwierigkeit bei der Lösung dieses Optimierungsproblems besteht darin, dass die Anzahl der falschen Segmentierungen $\overline{\boldsymbol{\pi}}$ exponentiell mit der Sequenzlänge zunimmt, wodurch das Lösen unter Berücksichtigung aller Nebenbedingungen nicht praktikabel ist. Es existieren jedoch iterative Verfahren, die effizient die sehr viel kleinere Menge der *aktiven* Nebenbedingungen finden. In meiner Arbeit beschäftigte ich mich auch damit, diese Verfahren weiter zu beschleunigen, ohne dabei an Vorhersagequalität einzubüßen [Zel10, und Referenzen darin].

Robustheit und Genauigkeit der HMSVM im Vergleich zum HMM

Im Rahmen der unten beschriebenen Anwendung auf die Suche nach transkribierten Bereichen untersuchte ich in meiner Dissertation empirisch den Einfluss der Trainingsmethode (HMSVM vs. HMM) auf die Effizienz des Lernalgorithmus, die Genauigkeit der Testvorhersagen und die Robustheit gegen künstlich eingefügtes Rauschen (Labelfehler).

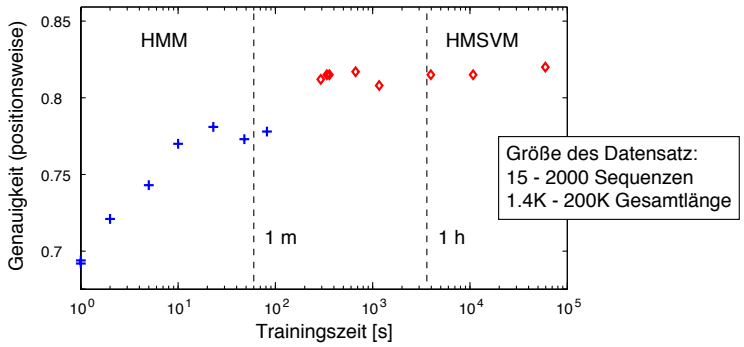

Abbildung 2: Vorhersagegenauigkeit des HMM und der HMSVM im Vergleich und in Abhängigkeit von der Trainingszeit. Beide Methoden wurden auf denselben Datensätzen trainiert bzw. ausgewertet. Die Grafik zeigt jeweils Mittelwerte aus drei verschiedenen Datensätzen gleicher Größe. (Entnommen aus [Zel10].)

Für die hier untersuchte Anwendung erreichte die HMSVM eine höhere Vorhersagegenauigkeit als das HMM und das bereits mit relativ wenigen Trainingsbeispielen (Abb. 2). Die diskriminativen Trainingsmethoden sind jedoch deutlich aufwändiger (Abb. 2), wohinge-

gen die Komplexität des Vorhersagealgorithmus bei beiden Methoden identisch ist: linear in der Sequenzlänge, was selbst für die Analyse großer Genome praktikabel ist.

Die HMSVM erwies sich nicht nur als genauer, sondern auch als robuster als das HMM. Ihre Vorhersagegenauigkeit verschlechterte sich nur um ca. 2 Prozentpunkte, selbst wenn 50% der Segmentlabel vertauscht wurden (die sich vertauschen ließen ohne ungültige Segmentierungen zu erzeugen). Eine ähnliche Minderung wurde für das HMM schon bei ca. 2% Rauschen beobachtet, bei 50% Rauschen verschlechterte sich die Genauigkeit um mehr als 10 Prozentpunkte [Zel10].

3 Anwendung der HMSVM auf biologische Probleme

Ein zentraler Teil meiner Arbeit bildet die Anwendung der HMSVM auf zwei wichtige biologische Fragestellungen: Zum einen die Suche nach Bereichen im Genom, in denen sich verschiedene Individuen einer Art durch *Genomsequenzvariationen* stark unterscheiden, im Unterschied zu genomischen Bereichen, die bei der überwiegenden Mehrheit innerhalb einer Art identisch sind. Zum anderen die *Transkript-Erkennung*, bei der es gilt, bestimmte Regionen von Genen (transkribierte Exons) von anderen genischen Segmenten (Introns) und der intergenischen Umgebung zu unterscheiden und die Struktur dieser Transkripte möglichst genau abzubilden (vgl. Abb. 1). In beiden Fällen wurden Tiling-Array-Daten analysiert, die im folgenden zunächst beschrieben werden.

Tiling-Arrays

Tiling-Arrays sind ein hochparalleles molekulares Verfahren, das es ermöglicht, an Millionen (im Falle von Resequencing-Arrays sogar Milliarden) von Messpunkten, die in einem regelmäßigen Raster das gesamte Genom repräsentieren, miniaturisierte DNA-Hybridisierungsexperimente durchzuführen. Diese DNA-Hybridisierung bezeichnet die spezifische Verbindung sequenzkomplementärer, einzelsträngiger DNA-Moleküle zu einer energetisch günstigeren Doppelhelixstruktur. Auf dem Tiling-Array sind einzelsträngige Sonden mit bekannter Sequenz (und Genomposition) aufgetragen, die an fluoreszenzmarkierte komplementäre DNA-Moleküle aus einem zu charakterisierenden, komplexen Zellextrakt binden (Abb. 3). Die Quantifizierung des Fluoreszenzsignals jeder einzelnen Sonde gibt Aufschluss über die Zusammensetzung des Zellextrakts. Besteht es aus der DNA eines Genoms, das Sequenzähnlichkeiten mit den Tiling-Sonden aufweist, wird ein stärkeres Hybridisierungssignal für identische Abschnitte der DNA erwartet als für solche, die sich in ihrer Sequenz unterscheiden. Besteht das Zellextrakt hingegen aus mRNA Transkripten, können diese nach Bindung an die Sonden ihrem genomischen Ursprung zugeordnet und quantifiziert werden – so lässt sich anschließend eine Aussage über die transkriptionelle Aktivität genomischer Bereiche machen (vgl. Abb. 3).

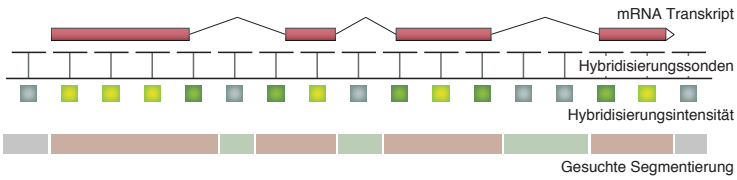

Abbildung 3: Illustration des Tiling-Arrays zur Erkennung transkribierter Bereiche. Auf einem Tiling-Array kann die transkriptionelle Aktivität des gesamten Genoms in regelmäßigen Intervallen mit Hybridisierungssonden gemessen werden. Die Intensität des Hybridisierungssignals ist proportional zur Menge komplementärer exonischer, fluoreszenzmarkierter Transkriptsequenzen (gelbgrün symbolisiert hier ein starkes Signal), und in intronischen und intergenischen Bereichen erwartet man nur ein schwaches Hintergrundsignal (grau). Transkriptkartierungsmethoden versuchen, ausgehend von den Hyrbidisierungssignalen, eine Segmentierung zu finden, die möglichst genau der Struktur der mRNA Transkripte entspricht. (Entnommen aus [Zel10].)

Detektion variabler Segmente im Genom

Nach Veröffentlichung der ersten menschlichen Referenzgenomsequenz [LLB$^+$01] verlagerte sich das Hauptaugenmerk der Genomforschung darauf zu verstehen, inwiefern sich individuelle Genome von dieser Referenz unterscheiden, welche Mutations- und Selektionsmechanismen dem zugrunde liegen und welche Konsequenzen sich aus Sequenzvarianten für den Organismus ergeben [Con10]. Sequenzvariation (Polymorphismen) im genomweit zu charakterisieren ist die Voraussetzung, um komplexe Merkmale eines Organismus' mit ursächlichen genetischen Varianten in Verbindung bringen zu können. Diese Verbindungen gewähren uns wiederum Einsichten in die molekularen Zusammenhänge vieler Krankheiten oder in evolutionäre Anpassungsprozesse. Die häufigste Klasse von Sequenzvarianten sind Einzelnukleotidänderungen (SNPs). Darüberhinaus existieren Insertionen und Deletionen, hochvariable Regionen und strukturelle Variationen wie z.B. Inversionen.

Heute hat die Revolution der DNA-Sequenziertechnologie [SJ08] mit drastisch erhöhtem Sequenzierdurchsatz und dramatischer Kostenreduktion die Sequenzierung individueller Genome im großen Stil möglich gemacht [Con10]. Zu Beginn meiner Doktorarbeit jedoch waren bestimmte Tiling-Arrays – sogenannte Resequencing-Arrays – noch die überlegene Technologie zur Erkennung von Polymorphismen. Diese Arrays ermöglichen im Prinzip die Detektion aller SNP-Varianten an allen nichtrepetitiven Positionen des Genoms [CST$^+$07]. Die resultierenden Messwerte weisen allerdings ein hohes Maß an zufälligem und systematischem Rauschen auf (vgl. Abb. 4), und die Erkennung von Sequenzvariationen über SNPs hinaus ist ungleich schwieriger [CST$^+$07].

In meiner Arbeit konnte ich zeigen, dass sich in diesen Daten dennoch polymorphe Genomregionen mit großer Genauigkeit erkennen lassen [ZCS$^+$08]. Mit Hilfe eines HMSVM-basierten Algorithmus fanden wir hunderttausende hochvariable Regionen (darunter $<3\%$ Falschpositive) im Pflanzenmodellorganismus *A. thaliana* (Ackerschmalwand). Diese beinhalten teils SNPs, teils Deletionen mit einigen wenigen bis zu tausenden von Nukleotiden. Aus diesen Resultaten entstand erstmals ein umfassendes, hochaufgelöstes Bild des Musters variabler Genomsequenzen in der Pflanze [CST$^+$07, ZCS$^+$08].

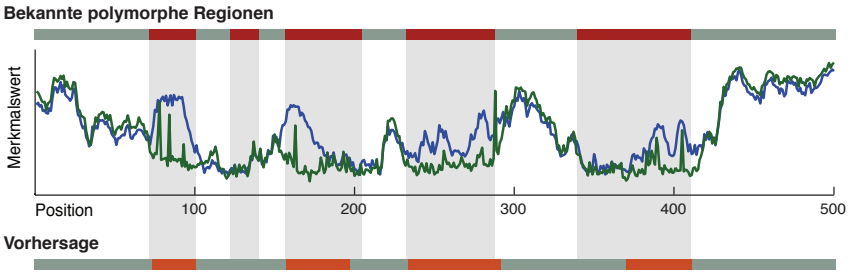

Abbildung 4: Vorhersage polymorpher genomischer Bereiche. Dargestellt ist ein Ausschnitt des Genoms mit einigen, bereits bekannten polymorphen Regionen (rote Balken oben und graue Schattierungen). Basierend auf Merkmalen, die von den Resequencing-Array-Signalen abgeleitet sind und auf Sequenzunterschiede im Individuum (grüne Linie, Mitte) relativ zur Referenzmessung (blau) hinweisen, wurde eine HMSVM-basierte Methode trainiert, die polymorphen Regionen wiederzuerkennen (hellrote Balken unten) und genomweit vorherzusagen. (Modifiziert nach [Zel10, ZCS$^+$08].)

Identifikation transkribierter Bereiche

Eine zweite Anwendung der HMSVM erfolgte mit dem Ziel, die Gesamtheit der Transkripte bestimmter Zellen, Gewebe oder Entwicklungsstadien (das *Transkriptom*) möglichst genau abzubilden. Das Problem ist eng mit einem Kernproblem der Bioinformatik, der Genvorhersage [SZZ$^+$09], verwandt. Im Gegensatz zur sequenzbasierten Genvorhersage verwendeten wir jedoch für die Transkriptsuche Tiling-Array Daten. Diese quantifizieren in einem regelmäßigen genomischen Raster die Menge an komplementären mRNA-Transkripten (Abb. 3) und liefern so ein Indikatorsignal für „Genaktivität“ (*Genexpression* genannt). Da Genexpression von Zell- und Organidentität sowie von Umwelteinflüssen abhängt, bildet eine genaue Transkriptomkartierung die Grundlage zur Rekonstruktion des Genregulationsnetzwerks, welches die Steuerung zellulärer Prozesse, z.B. der Zelldifferenzierung bis hin zur Organ- und Organismenentwicklung, beschreibt. Aufgrund dieser grundlegenden Bedeutung für Genetik und Entwicklungsbiologie ist ein besseres Verständnis des Transkriptoms von höchstem Interesse [GBD$^+$10].

Im Rahmen meiner Arbeit entwickelte ich die HMSVM-Segmentierung zu einem Verfahren zur *Transkriptsuche und -Kartierung* mit Hilfe von Tiling-Array-Daten (mSTAD) [ZHL$^+$08]. Angewandt auf Daten für die Modellorganismen *A. thaliana* und *C. elegans* (ein kleiner Fadenwurm mit großer Bedeutung für die Molekularbiologie), zeigte sich, dass mSTAD eine erheblich genauere Transkriptkartierung ermöglicht als bisherige Verfahren [KCK$^+$04, LZH$^+$08, ZHW$^+$09, SZW$^+$11, GLVN$^+$10] und dass der diskriminative HMSVM-Trainingsalgorithmus dem generativen HMM überlegen ist (Abb. 5) [Zel10].

Trotz der verbesserten Übereinstimmung zwischen mSTADs Vorhersagen und experimentell charakterisierten Transkripten identifizierten wir in genomweiten Analysen tausende neue Transkripte, die ungeachtet großer vorhergehender Annotationsprojekte zuvor unentdeckt geblieben waren. Von einer Auswahl der Vorhersagen konnten $>$75% experimentell bestätigt werden [LZH$^+$08]. Unsere Resultate trugen damit erheblich zur Vervollständigung der Transkriptkartierung in diesen beiden Modellorganismen bei. Unse-

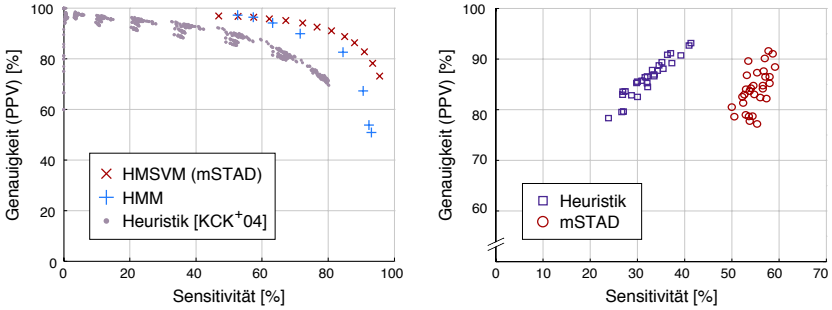

Abbildung 5: Vergleich der Vorhersagegenauigkeit verschiener Transkripterkennungsverfahren. Die in meiner Arbeit entwickelte HMSVM-basierte Methode, mSTAD, wird hier verglichen gegen ein HMM und eine weitverbreitete Heuristik [KCK+04], die eine Sliding-Window-Technik benutzt und auch für modMine eingesetzt wurde [GLVN+10]. Die linke Abbildung zeigt Vorhersagen für verschiedene Konfigurationsparameter (auf *A. thaliana* Daten); die rechte Vorhersagen für 25 verschiedene modENCODE-Datensätze (*C. elegans*) mit einem festen Parameterset [GLVN+10, SZW+11]. (PPV: Positive Predictive Value; entnommen aus [Zel10, SZW+11].)

re Analyse verschiedener Organe, Zelltypen, Entwicklungsstadien und Umwelteinflüsse [LZH+08, ZHW+09, SZW+11] zeichnete ein dynamisches Bild des Transkriptoms (und in dieser Hinsicht gehen unsere Ergebnisse weit darüber hinaus, was man mit Genvorhersage erreichen kann). Die Mehrzahl der neu entdeckten Transkripte wies spezialisierte Expressionsmuster auf und war nur in einer kleinen Teilmenge der untersuchten Gewebe oder Bedingungen detektierbar. Das liefert einerseits eine mögliche Erklärung, warum diese in früheren Studien, die meist den Gesamtorganismus bei Standardbedingungen untersuchten, nicht gefunden wurden. Andererseits legt diese Beobachtung nahe, dass wir erst mit einer Untersuchung der Transkriptome aller Zelltypen eines Organismus ein vollständiges Bild von der Gesamtzahl seiner Gene erhalten werden [SZW+11].

Viele der neu entdeckten Transkripte dienen vermutlich nicht direkt der Produktion von Proteinen (wir bezeichnen sie daher als „nichtkodierend“). Die Funktionen nichtkodierender Transkripte im einzelnen aufzuklären ist seit kurzem Gegenstand intensiver biologischer Forschung; erste spektakuläre Durchbrüche lassen bereits eine Vielfalt regulatorischer Funktionen erkennen [WSS09].

4 Diskussion und Ausblick

Der Hauptbeitrag meiner Dissertation bestand darin, Methoden des Maschinellen Lernens weiterzuentwickeln um Daten aus Hochdurchsatzexperimenten aus dem Bereich der Genom- und Transkriptomforschung zu analysieren. Besonders konzentrierte ich mich dabei auf Segmentierungsmethoden, da sich viele wichtige Fragestellungen der Bioinformatik als Segmentierungsprobleme formalisieren lassen. Im folgenden werde ich kurz diskutieren, warum sich die HMSVM als ideale Technik erwiesen hat und warum ich darüberhinaus großes Potential für eine breite Anwendung sehe.

Zunächst hat der Modellierungsansatz der HMSVM große Ähnlichkeit mit dem in der Bio-

informatik weit verbreiteten HMM, was ihr ein extrem *vielfältiges Einsatzgebiet* eröffnet. Im Anschluss an meine Dissertation entwickelte ich die HMSVM-basierte Transkripterkennung weiter, um damit auch Transkriptom-Sequenzierungsdaten, die mit neuesten DNA-Sequenziertechniken [WGS09] generiert wurden, analysieren zu können, da diese in naher Zukunft Tiling-Arrays als Standardtechnologie ablösen werden. Weitere Anwendungsmöglichkeiten umfassen Genvorhersage, Sequenz-Alignment, Protein-Homologiesuche, Protein-Domänen-Annotation, RNA-Strukturvorhersage und viele mehr.

Das diskrimative HMSVM-Training resultiert in einer *hohen Vorhersagegenauigkeit*, die für die biologische Datenauswertung den großen Vorteil hat, dass zeit- und kostenintensive Folgeexperimente seltener durch falsche Vorhersagen fehlgeleitet werden. Für die Transkripterkennung zeigte meine Arbeit im Vergleich zum HMM und der meistgenutzten heuristischen Methode signifikant verbesserte Genauigkeit.

Dass die HMSVM sich durch große *Robustheit gegenüber Messfehlern und Rauschen* auszeichnet, macht sie zur idealen Methode für die Analyse biologischer Datensätze.

Schließlich ist die *Integration heterogener Merkmalstypen* einfach möglich, da im Gegensatz zum HMM (und anderen probabilistischen Methoden) bei der Modellierung nur schwache Annahmen gemacht werden. In meiner Arbeit untersuchte ich beispielsweise wie sich sequenzbasierte und Tiling-Array-basierte Merkmale kombinieren lassen um das Transkriptsucheverfahren weiter zu verbessern [Zel10]. Allgemein ist diese einfache Integrationsmöglichkeit ein entscheidender Vorteil, da für Modellorganismen genomweite, experimentelle Datensätze in zuvor ungekannter Vielfalt öffentlich zugänglich sind. Diese beschreiben viele Aspekte der Genomorganisation, Transkription und ihrer Regulation [GLVN$^+$10, u.a.]. Es zeichnet sich bereits ab, dass Machinelle Lernmethoden zur Integration dieser vielfältigen Daten in steigendem Maße zur Aufklärung der zugrundeliegenden komplexen biologischen Prozesse und Systeme beitragen werden [GLVN$^+$10, BCG$^+$10].

Danksagung

Herzlich danke ich den Betreuern meiner Arbeit, Gunnar Rätsch und Detlef Weigel, und den Mitgliedern ihrer Gruppen.

Literatur

[ATH03] Y. Altun, I. Tsochantaridis und T. Hofmann. Hidden Markov Support Vector Machines. *Proceedings of the ICML*, Seiten 3–10, 2003.

[BCG$^+$10] Y. Barash, J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe und B. J. Frey. Deciphering the splicing code. *Nature*, 465:53–59, 2010.

[Con10] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–73, 2010.

[CST$^+$07] R. M. Clark, G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, P. Shinn, N. Warthmann, T. T. Hu, G. Fu, D. A. Hinds, *et al*. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, 317:338–42, 2007.

[DEKM98] R. Durbin, S. Eddy, A. Krogh und G. Mitchison. *Biological Sequence Analysis: Probabilistic models of protein and nucleic acids*. Cambridge University Press, 7. Auflage, 1998.

[GBD$^+$10] B. R. Graveley, J. W. Brooks, A. N. Carlson, M. O. Duff, J. M. Landolin, L. Yang, C. G. Artieri, M. J. van Baren, N. Boley, B. W. Booth, *et al*. The developmental transcriptome of Drosophila melanogaster. *Nature*, 471:473–479, 2010.

[GLVN+10] M. B. Gerstein, Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T Liu, K. Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, *et al*. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, 330:1775–87, 2010.

[KCK+04] D. Kampa, J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, *et al*. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Research*, 14:331–42, 2004.

[LLB+01] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, *et al*. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[LZH+08] S. Laubinger, G. Zeller, S. R. Henz, T. Sachsenberg, C. K. Widmer, Naouar N, M. Vuylsteke, B. Schölkopf, G. Rätsch und D. Weigel. At-TAX: A whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. *Genome Biology*, 9:R112, 2008.

[SJ08] J. Shendure und H. Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 10:1135–1145, 2008.

[SS02] B. Schölkopf und A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[SZW+11] W. C. Spencer, G. Zeller, J. D. Watson, S. R. Henz, K. L. Watkins, R. D. McWhirter, S. Petersen, V. T. Sreedharan, C. Widmer, J. Jo, *et al*. A spatial and temporal map of C. elegans gene expression. *Genome Research*, 21:325–41, 2011.

[SZZ+09] G. Schweikert, A. Zien, G. Zeller, J. Behr, C. Dieterich, C. Ong, P. Philips, F. De Bona, L. Hartmann, A. Bohlen, *et al*. mGene: Accurate SVM-based gene finding with an application to nematode genomes. *Genome Research*, 19:2133–43, 2009.

[Vap95] V. N. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 1995.

[WGS09] Z. Wang, M. Gerstein und M. Snyder. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63, 2009.

[WSS09] J. E. Wilusz, H. Sunwoo und D. L. Spector. Long noncoding RNAs: functional surprises from the RNA world. *Genes and Development*, 23:1494–1504, 2009.

[ZCS+08] G. Zeller, R. M. Clark, K. Schneeberger, A. Bohlen, D. Weigel und G. Rätsch. Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. *Genome Research*, 18:918–29, 2008.

[Zel10] G. Zeller. *Machine Learning Algorithms for the Analysis of Data from Whole-Genome Tiling Microarrays*. Dissertation, Universität Tübingen, 2010.

[ZHL+08] G. Zeller, S. R. Henz, S. Laubinger, D. Weigel und G. Rätsch. Transcript normalization and segmentation of tiling array data. *Pacific Symposium on Biocomputing*, Seiten 527–38, 2008.

[ZHW+09] G. Zeller, S. Henz, C. Widmer, T. Sachsenberg, G. Rätsch, D. Weigel und S. Laubinger. Stress-induced changes in the Arabidopsis thaliana transcriptome analyzed using whole genome tiling arrays. *The Plant Journal*, 58:1068–82, 2009.

Georg Zeller, geboren 1979, studierte Bioinformatik, Informatik und Molekularbiologie in Tübingen und Uppsala (Schweden). Nach dem Diplom in Informatik / Bioinformatik (2006) verfasste er seine Dissertation am Friedrich-Miescher-Laboratorium der Max-Planck-Gesellschaft in Dr. Gunnar Rätschs Gruppe und am Max-Planck-Institut für Entwicklungsbiologie in Prof. Dr. Detlef Weigels Abteilung in Tübingen. Seit 2010 ist er am Europäischen Laboratorium für Molekularbiologie in Heidelberg bei Dr. Peer Bork tätig, wo er seine wissenschaftlichen Interessen im Bereich der Bioinformatik und Chemoinformatik weiter verfolgt.