Aus der Medizinische Klinik mit Schwerpunkt Hämatologie, Onkologie
und Tumorimmunologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Charakterisierung der EpCAM-reaktiven Immunität in Karzinompatienten

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Dr. rer. nat. Oliver Schmetzer

aus Bayreuth

Datum der Promotion: 25. Oktober 2013

Inhaltsverzeichnis

ORIGINALARBEITEN, DIE DIESER PROMOTION ZUGRUNDE LIEGEN

Publikation 1

Schmetzer O, Moldenhauer G, Riesenberg R, Pires JR, Schlag P, Pezzutto A.

Quality of recombinant protein determines the amount of autoreactivity detected against the tumor-associated epithelial cell adhesion molecule antigen: low frequency of antibodies against the natural protein.

J Immunol. 2005 Jan 15; 174(2): 942–952.

Impact factor: **5,788** (2011 Journal Citation Reports)

Publikation 2

Schmetzer O, Moldenhauer G, Nicolaou A, Schlag P, Riesenberg R, Pezzutto A.

Detection of circulating tumor-associated antigen depends on the domains recognized by the monoclonal antibodies used: N-terminal trimmed EpCAM-levels are much higher than untrimmed forms.

Immunol Lett. 2012 Apr 30; 143(2):184–192. Epub 2012 Feb 22.

Impact factor: **2,526** (Fünfjahresmittelwert, 2012 Journal Citation Reports)

Publikation 3

Riedesel H, Kolbeck B, Schmetzer O, Knapp EW.

Peptide binding at class I major histocompatibility complex scored with linear functions and support vector machines.

Genome Inform. 2004; 15(1):198–212.

Impact factor not determined – 2012 Still Computing

Publikation 4

Salmon D, do Aido-Machado R, Diehl A, Leidert M, Schmetzer O, de A Lima AP, Scharfstein J, Oschkinat H, Pires JR.

Solution structure and backbone dynamics of the Trypanosoma cruzi cysteine protease inhibitor chagasin.

J Mol Biol. 2006 Apr 14;357(5):1511-21

Impact factor: **4,001** (2011 Journal Citation Reports)

ABSTRACT

The human epithelial cell adhesion molecule (EpCAM) is expressed on normal epithelial cells but overexpressed in most carcinomas. EpCAM-targeted immunotherapy has been reported in clinical studies and is currently approved by using EpCAM-CD3-bispecific antibodies.
Due to the high side effects and so far limited clinical efficiency the aim of this study was to analyze the preexisting humoral response including anti-idiotypic antibodies and soluble EpCAM concentrations with a special emphasis on the purity and folding of the used protein.
Anti-EpCAM positive sera were found in a much lower percentage as previously published which was due to reactivity against denatured protein. The importance of correctly folded protein could be proven by NMR- and CD-analysis in the EpCAM-ELISA, but also by structure determination with other proteins (Chagasin).
In direct comparison, Tetanus toxoid reactive antibodies of the IgG isotype were present in 1000 fold higher concentration as compared to EpCAM-reactive antibodies. In contrast IgA isotype antibodies showed a higher concentration against EpCAM.
We developed an ELISA-method to quantify the circulating soluble EpCAM protein in sera. High concentrations could be identified in 471 human sera with a mean concentration of 2 µg/ml. EpCAM reactive IgG antibodies were tested in the same sera but could only be detected in sera with less than 1 µg/ml circulating EpCAM.
Therapeutic antibodies against EpCAM have been developed and used in clinical trials since more than 30 years. To lower side effects, the presence of high levels of circulating EpCAM and of anti-idiotype antibodies must be kept in mind.

249 words

ZUSAMMENFASSUNG

Die vorliegende Arbeit beschäftigt sich mit der humoralen Immunität in Karzinompatienten mittels Messung und Charakterisierung von Antikörpern gegen das epitheliale Zelladhäsionsmolekül (EpCAM) sowie mit der Messung von löslichem EpCAM in Patientenseren. Methoden zur Messung der Konzentration von Anti-EpCAM-Antikörpern, zu deren Subtypisierung und zur Messung von löslichem EpCAM wurde etabliert. Dabei wurden Methoden optimiert, die lösliches Protein in hoher Qualität und nativer Faltung reinigen, welche auch für die NMR-Strukturbestimmung geeignet sind. Hohe lösliche EpCAM-Werte wurden als möglicher Grund für die hohe Toleranz in Patienten mit fehlender oder stark reduzierter Antikörperantwort identifiziert. Um die Messung der zellulären Immunantwort zu verbessern, wurden neue Algorithmen zur Epitopvorhersage entwickelt.

Zu Beginn der Studien wurde rekombinantes EpCAM in Insektenzellen hergestellt, um vollständig glykosyliertes und nativ gefaltetes Antigen zu erhalten. Die Expressionsvektoren wurden im Gegensatz zu anderen Studien so gewählt, dass eine Sekretion des Proteins stattfindet, sodass im Golgi-Apparat fehlerhaft gefaltetes Protein und unvollständige Proteinfragmente entfernt werden. Weitere Reinigungsschritte des Proteins (v.a. Ionenaustauscher- und hydrophobe Interaktionschromatografie) wurden eingesetzt, um Reste nicht-nativen Proteins zu entfernen.

Sekundärstruktur (mittels zirkularer Dichroismusspektroskopie) und Tertiärstruktur (mittels NMR-Spektroskopie) des Proteins wurden mit kommerziell erhältlichem und in anderen Studien benutztem EpCAM-Protein verglichen. Dabei zeigte sich, dass durch das Expressionssystem (E. coli, Seidenspinnerraupen) und die Reinigung des kommerziell verfügbaren Proteins, das auch in anderen Gruppen zu hohen positiven Werten geführt hatte, nicht-glykosyliertes, fehlerhaft gefaltetes Protein angereichert wurde. Dies wurde auch bei einer NMR-Studie mit dem Chagasin-Molekül validiert. Weder die hohen Prozentzahlen positiver Patienten in den ELISAs (in der Literatur um die 50%) noch andere Ergebnisse der in der Literatur oft verwendeten durchflusszytometrischen Shift-Assays (in der Literatur um die 80% Anti-EpCAM positive Seren) konnten reproduziert werden, die Spezifität dieses Assays ist aber gering. Auch in direkten Experimenten (Immunpräzipitation, Western Blots) konnten nur in einem sehr geringen Patientenanteil tatsächlich Anti-EpCAM-Antikörper nachgewiesen werden. Im direkten Vergleich wurde eine deutlich reduzierte Zahl positiver Proben gefunden (44 von 500 getesteten Seren). Mit dem von anderen Gruppen benutzten, denaturierten Protein wurden 12-mal mehr positive Seren mit teilweise 1000-fach höheren

Anti-EpCAM-Mengen gemessen. Als Vergleich nutzten wir eine Antikörpermessung gegen Tetanustoxoid, die im Mittel eine 1000-fach höhere Konzentration aufwies. Auch konnten wir nachweisen, dass eine hohe Reaktivität gegen bovines Serumalbumin vorliegt, was in früheren Studien einen Teil der hohen Werte erklären lässt.
In der Literatur war mehrfach beschrieben worden, dass beinahe alle Epithelzellen EpCAM exprimieren und hohe Serumkonzentrationen von Protein häufig mit der humoralen Toleranz korrelieren. Um die Ursache für die geringe Immunreaktion zu finden, wurde daher das von der Zelloberfläche abgeschnittene Protein gemessen. Mit zwei in Kooperation (Dr. Gerd Moldenhauer, DKFZ Heidelberg) hergestellten, neuen monoklonalen Antikörpern wurde dann ein ELISA-System etabliert. Hierzu wurden auch polyklonale Seren und andere gut charakterisierte monoklonale Antikörper gegen EpCAM getestet. Ein fluorogener ELISA wurde etabliert und optimiert, der über mehrere log-Stufen eine Messung der Antikörper ermöglicht und eine 1 bis 2 log-Stufen höhere Sensitivität als chromogene ELISAs besitzt. Wir fanden hohe Konzentrationen von EpCAM, die negativ mit der Antikörperantwort korrelierten, wie es auch schon für andere Proteine beschrieben wurde.

EINFÜHRUNG

Das epitheliale Zellädhäsionsmolekül (EpCAM) war eines der ersten tumorassoziierten Antigene, das beschrieben wurde. 1979 erfolgte die Beschreibung durch einen monoklonalen Antikörper, molekularbiologisch identifiziert und charakterisiert wurde es 1986 [1]. Nachdem die hohe Immunogenität und Spezifität beschrieben worden waren, wurden über 100 klinische Studien bei Krebspatienten mit EpCAM als Target-Molekül durchgeführt. Vor allem monoklonale Antikörper und Anti-Idiotyp-Präparationen wurden neben traditionellen und neuartigen Vakzinen getestet [2]. Der Anti-EpCAM-Antikörper CO17-1A (Panorex®) wurde zugelassen, verlor die Zulassung aber, nachdem die Behandlung in der adjuvanten Situation zu schlechteren Ergebnissen als die Zytostatika geführt hatte.

Die Vielzahl der klinischen Studien bei mäßigen Ergebnissen lässt sich einerseits durch die hohe Immunogenität erklären, die vor allem mit einer ausgesprochen starken humoralen Immunantwort beschrieben wurde. Allerdings wurde nie ein validierter Test zur Messung der humoralen Immunantworten entwickelt und in den meisten Studien wurden unspezifische durchflusszytometrische Bindungen an Tumorzelllinien als Anti-EpCAM-Immunreaktion gewertet. Andererseits wurde angenommen, dass EpCAM auf Tumorzellen im Vergleich zu gesundem Gewebe stark überexprimiert wird. Neue Studien zeigen, dass dies eher ein Effekt einzelner Antikörper war, die zu den Färbungen benutzt wurden, und die Expression von EpCAM im Gegensatz zu Tumorgewebe vor allem in gesunden Stammzellen der Leber, des Pankreas, der Nieren u.a. erhöht vorliegt [3]. Dennoch verschlechtert eine erhöhte EpCAM-Expression des Tumors die Prognose, was aktuell eher als Zeichen einer aggressiveren, von früheren Stammzellen abgeleiteten Tumorvariante gilt. Vor einigen Jahren wurde auch gezeigt, das EpCAM direkt als Signalmolekül wirkt und die Tumorproliferation direkt fördern kann [4]. Es reguliert in Tumorzellen vor allem die Expression von micro-RNAs und damit Tumorzellproliferation und -invasion [5]. EpCAM-knock-out-Mäuse sterben *in utero* aufgrund abnormaler Plazentaentwicklung [6]. Das Protein wird in allen Organen exprimiert, im konditionalen Knock-out-Modell konnte eine essenzielle Funktion bei der Induktion von Immunantworten, vor allem bei der Aktivierung und Migration der Langerhanszellen gezeigt werden [7].

Ziele dieser Arbeit waren, die hohe Immunogenität des EpCAMs kritisch zu prüfen, die beschriebenen humoralen Antworten weiter zu charakterisieren, EpCAM als Tumormarker zu untersuchen und die Identifikation von T-Zell-Epitopen des EpCAM durch neue Techniken zu ermöglichen.

METHODIK

Die detaillierte technische Beschreibung findet sich in den Nachdrucken der Veröffentlichungen. Einige Punkte waren für die erfolgreiche Durchführung der Arbeit jedoch essenziell: Zur Expression des rekombinanten EpCAMs wurde die Drosophilazelllinie „Schneider S2“ benutzt. Diese kann in proteinfreiem Medium kultiviert werden, wodurch eine Verunreinigung des rekombinanten Proteins geringer wird. Wichtiger dagegen ist die Verwendung eines Sekretionspeptids aus der Honigbiene, sodass nur vollständig gefaltetes und glykosyliertes EpCAM im Überstand enthalten ist. Die rekombinanten Varianten des EpCAMs in anderen Studien wurden in E. coli, Seidenspinnerlarven oder mittels Baculovirusinfektion hergestellt. All diese Techniken führen nicht zu korrekt gefaltetem Protein, da die Zellen zur Reinigung lysiert werden und so fehlgefaltetes, unvollständiges und nicht komplett glykosyliertes Protein nicht im Golgi abgesondert werden kann. Dies konnte insbesondere in der vierten Publikation mittels einer NMR Strukturbestimmung gezeigt werden. Nicht-glykosyliertes, intrazellulär vorliegendes Protein kann nach extensiver Reinigung nativ aus E. Coli verwendet werden, bei extrazellulären oder Transmembranproteinen führt eine Expression in E. Coli zu stark fehlgefaltetem Protein wie in den Publikationen 1 und 2 beschrieben.

ERGEBNISSE UND DISKUSSION

Initial wurden Patientenseren mit den traditionellen durchflusszytometrischen Shift-Assays auf Anti-EpCAM-IgG ohne Subgruppentypisierung getestet. Dieses Testsystem wurde in den ersten Publikationen zu EpCAM benutzt, an Kolonkarzinomlinien bindende Antikörper wurden nachgewiesen und dies wurde als Beweis der Immunogenität des EpCAMs angesehen [8-10]. Jedoch konnten in dieser Arbeit nur schwache Signale erhalten werden und die Spezifität der gemessenen Antikörper konnte nicht durch Kompetitionsexperimente validiert werden. In der ersten Veröffentlichungen dieser Arbeit wurde lediglich eine leicht erhöhte mittlere Konzentration in einzelnen Patienten an Anti-EpCAM-Antikörpern beschrieben. Im direkten Vergleich zu Tetanustoxoid-reaktiven IgG ist die Konzentration des EpCAM-reaktiven IgGs um den Faktor 1000 geringer. Um den Messbereich dieser geringen Menge zu erfassen, wurden fluorogene ELISA-Systeme mit rekombinantem EpCAM und neuen Anti-EpCAM-Antikörpern etabliert, da fluorogene ELISA oft 10- bis 100-fach sensitiver sind als photometrische [11-13]. Antikörper-Epitope von Anti-EpCAM-IgG in Gesunden waren kongruent zu unseren Ergebnissen identifiziert worden, allerdings war in dieser Studie keine Konzentrationsbestimmung erfolgt [14]. Tatsächlich gelang es sowohl fluorogene ELISAs

zum Nachweis von Anti-EpCAM-Antikörpern (Publikation 1) als auch zum direkten Nachweis von löslichem EpCAM (Publikation 2) mit dem rekombinanten EpCAM aus den Insektenzellen und den damit erzeugten neuen monoklonalen Antikörpern mit guter Nachweisgrenze (pg/ml Bereich) zu etablieren. Unsere Ergebnisse einer lediglich diskret erhöhten löslichen EpCAM-Konzentration in Karzinompatienten im Vergleich zu Gesunden wurden in einer aktuellen Studie von 2012 einer anderen Gruppe bestätigt [15]. In dieser Studie zeigten Messungen bei der Bestimmung von EpCAM als potenzieller Tumormarker eine höhere Sensitivität anderer löslicher Tumormarker (z.B. CA 19-9) im Vergleich zu EpCAM. Allerdings unterschieden sich die Ergebnisse unseres ELISAs, vor allem die mittlere Konzentration des löslichen EpCAMs und die Spezifität des Testsystems für Karzinomerkrankungen, stark von einer älteren Publikation aus 2002 [16]. In der zweiten Publikation dieser Arbeit konnten die initial diskrepanten Ergebnisse zu dem später auch kommerziellen ELISA dieser Arbeit aus 2002 im direkten Vergleich erklärt werden: Unterschiedliche Standardproteine waren verwendet und unterschiedliche Domänen des EpCAMs detektiert worden, sodass verkürzte Formen im kommerziellen ELISA nicht detektiert worden waren. Dies führte bei dem von uns entwickelten ELISA zu um zwei log-Stufen höheren absoluten Konzentrationen.

Lösliches EpCAM wurde bis 2008/2009 als möglicher Tumormarker bei bestimmten Karzinompatienten identifiziert. Erst ab 2009 wurde physiologisch die Spaltung der extrazellulären Domäne unter physiologischen Bedingungen gezeigt [4, 17, 18]. Lösliches EpCAM wurde dann auch im Serum identifiziert, was unsere Daten weiter bestätigte [19]. Allerdings weichen die Konzentrationsangaben auch hier deutlich ab. Wie in der zweiten Publikation dieser Arbeit gezeigt, führten wieder die unterschiedlichen Antikörper zur Detektion unterschiedlicher löslicher EpCAM-Formen, wobei in der publizierten Studie nur N-terminal intaktes EpCAM gemessen wurde. Dieses macht nur einen geringen Anteil des EpCAMs im Serum aus, das meiste Protein ist N-terminal verkürzt und wird so von den publizierten ELISAs nicht miterfasst.

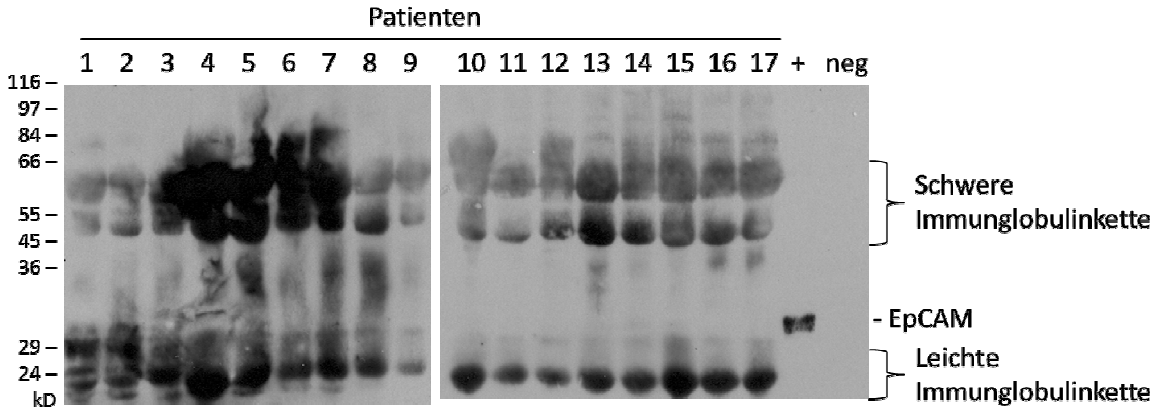

Abbildung 1: Westernblot mit biotinyliertem CO17-1A. 17 Patientenseren wurden aufgetragen, die zuvor besonders hohe oder niedrige Werte an EpCAM-Antikörpern im ELISA zeigten. +: 50 ng rekombinantes EpCAM-Protein wurde als Positivkontrolle eingesetzt, in der Spur daneben eine Negativkontrolle mit einem anderen unspezifischen Protein (LAIR).

Wir testeten die Patientenseren auch auf verschiedene Anti-Idiotyp-Antikörper. Gegen CO17-1A konnten mehrere Immunglobulin-Isotypen als Idiotyp-Antikörper identifiziert werden (Abb. 1). Gegen HEA125, einen anderen Anti-EpCAM-Antikörper war lediglich ein Isotyp vorhanden. Bei anderen Anti-EpCAM-Antikörperklonen konnten keine Idiotypen festgestellt werden. Daraus lässt sich folgern, dass nur bestimmte Teile des EpCAM-Proteins zu molekularem Mimikry mittels Anti-Idiotyp-Antikörper führen und dass diese im Falle von CO17-1A aufgrund des Molekulargewichts der schweren Kette sowohl vom IgG1 oder IgG2 aber auch vom IgG3-Typ sein müssen. Die von uns nachgewiesenen Anti-Idiotyp-Antikörper korrelierten nicht mit dem Anti-EpCAM-Antikörper-Titer. In allen getesteten Proben konnten diese Anti-Idiotypen gegen CO17-1A gefunden werden, auch in den am stärkstem Anti-EpCAM-positiven und -negativen Seren. Somit konnten wir die Anti-CO17-1A-Antikörper bereits in nicht immunologisch behandelten Patienten nachweisen. Dies stimmt mit den Ergebnissen von anderen Gruppen überein, die auch EpCAM-reaktive Antikörper in gesunden Probanden beschrieben hatten [14].

In früheren Studien war ein Netzwerk von Anti-EpCAM-Idiotyp-Antikörpern beschrieben und extensiv untersucht worden [8, 9, 20-28]: Studien mit direkten Anti-EpCAM-Antikörpern (Ab1) wurden im Sinne einer passiven Immunisierung vor allem mit Anti-CO17-1A-Antikörper erstmals 1986 in 142 Patienten durchgeführt, jedoch wurden die pharmakokinetischen Aspekte bei der Applikation von Panorex® nicht berücksichtigt [27]. Um eine aktive Immunisierung und damit höhere Ansprechraten zu erreichen, wurden dann klinische Studien mit Anti-Idiotyp-Antikörper (Ab2) durchgeführt. Vor allem CO17-1a wurde zur Induktion von Anti-Idiotyp-Antikörpern (Ab2) in Ziegen zur Gewinnung von Seren für

klinische Studien getestet [26]. Patienten wurden mit diesen polyklonalen Anti-Idiotyp-Antikörpern (Ab2) behandelt und in allen 30 Patienten konnten Anti-Anti-Idiotyp-Antikörper (Ab3) und damit eine erfolgreiche aktive Immunisierung nachgewiesen werden. Die Ab2 sollten hierbei als molekulares Mimikry des EpCAM wirken und der Nachweis der Ab3 in allen Patienten wurde als erfolgreiche Induktion einer Anti-EpCAM-Antwort gewertet. Tatsächlich konnte in dieser Studie in 6 von 30 Patienten eine partielle klinische Remission und in weiteren 7 Patienten eine „stable disease“ erreicht werden. Von den 13 Patienten, die auf das Ziegenserum ansprachen, wurden 4 ausschließlich mit Anti-Idiotypen behandelt, die anderen erhielten zusätzlich eine zytostatische Therapie. Grundlage für die speziesübergreifende Aktivität und Transferierbarkeit des Idiotyp-Netzwerks im Sinne einer passiven und auch aktiven Immunisierung waren Studien, die gezeigt hatten, dass polyklonale Seren sowie monoklonale Antikörper dieses molekulare Mimikry speziesunabhängig übertragen konnten, unabhängig davon, ob Kaninchen, Mäuse oder Ziegen immunisiert worden waren [28]. Die Existenz dieser Anti-CO17-1A-Antikörper konnte in dieser Arbeit also verifiziert werden. Allerdings wurde in den klinischen Studien kein Anstieg des Titers der spezifischen Anti-Idiotyp-Antikörper gemessen, sodass fraglich bleibt, inwieweit das offensichtlich natürliche Anti-Idiotyp-Netzwerk gegen CO17-1A tatsächlich durch die Gabe von CO17-1A oder Seren beeinflusst wurde.

Aufgrund des vorhandenen löslichen EpCAMs in hoher Konzentration und des Idiotyp-Netzwerks, welche lediglich die Nebenwirkung erhöhen, sollte bei einer Immuntherapie zuerst der mögliche Absorptionseffekt überprüft werden.

Die dritte Publikation dieser Arbeit beschreibt ein neues mathematisches Modell zur Vorhersage von T-Zell-Epitopen. Vor allem der Klassenwechsel zu Anti-EpCAM-IgG in Gesunden weist stark auf die Existenz von EpCAM-reaktiven T-Zellen hin, die dazu bei nicht-repetitiven Antigenen benötigt werden. Auf Basis einer Sequenzdatenbank mit bekannten T-Zell-Epitopen und Kontrollpeptiden konnten wir zeigen, dass die Vorhersage am besten durch Lösen eines linearen Gleichungsmodells erfolgt. Andere Modelle, wie die zurzeit in Vorhersagen oft genutzten Homologie- oder Konsensussequenzmodelle, sowie neuere, auf neuronalen Netzwerken beruhende Modelle waren deutlich unterlegen.

LITERATURVERZEICHNIS

[1] Herlyn M, Steplewski, Z., Herlyn, D., and Koprowski, H.: Colorectal carcinoma-specific antigen: Detection by means of monoclonal antibodies. Proc Natl Acad Sci U S A 1979;76:1438.

[2] Birebent B, Somasundaram, R., Purev, E., Li, W., Mitchell, E., Hoey, D., Bloom, E., Mastrangelo, M., Maguire, H., Harris, D. T., Staib, L., Braumuller, H., Lesser, C., Kuttner, N., Beger, H-G., and Herlyn, D.: Anti-idiotypic antibody and recombinant antigen vaccines in colorectal cancer patients. Critical Reviews in Oncology/Hematology 2001;39:107.

[3] Yamashita T, Budhu A, Forgues M, Wang XW: Activation of hepatic stem cell marker EpCAM by Wnt-beta-catenin signaling in hepatocellular carcinoma. Cancer Res 2007;67:10831.

[4] Maetzel D, Denzel S, Mack B, Canis M, Went P, Benk Met al. : Nuclear signalling by tumour-associated antigen EpCAM. Nat Cell Biol 2009;11:162.

[5] Kandalam MM, Beta M, Maheswari UK, Swaminathan S, Krishnakumar S: Oncogenic microRNA 17-92 cluster is regulated by epithelial cell adhesion molecule and could be a potential therapeutic target in retinoblastoma. Mol Vis 2012;18:2279.

[6] Nagao K, Zhu J, Heneghan MB, Hanson JC, Morasso MI, Tessarollo Let al. : Abnormal placental development and early embryonic lethality in EpCAM-null mice. PLoS One 2009;4:e8543.

[7] Gaiser MR, Lammermann T, Feng X, Igyarto BZ, Kaplan DH, Tessarollo Let al. : Cancer-associated epithelial cell adhesion molecule (EpCAM; CD326) enables epidermal Langerhans cell motility and migration in vivo. Proc Natl Acad Sci U S A 2012;109:E889.

[8] Wettendorff M, Iliopoulos, D., Tempero, M., Kay, D., DeFreitas, E., Koprowski, H., and Herlyn, D.: Idiotypic cascades in cancer patients treated with monoclonal antibody CO17-1A. Proc Natl Acad Sci U S A 1989;86:3787.

[9] Herlyn D, Ross, A.H., Iliopoulos, D., and Koprowski, H.: Induction of specific immunity to human colon carcinoma by anti-idiotypic antibodies to monoclonal antibody CO17-1A. Eur J Immunol 1987;17:1649.

[10] Ross AH, Lubeck M, Steplewski Z, Koprowski H: Identification and characterization of the CO17-1A carcinoma-associated antigen. Hybridoma 1986;5 Suppl 1:S21.

[11] Chargelegue D, O'Toole CM: Development of a sensitive ELISA for HIV-1 p24 antigen using a fluorogenic substrate for monitoring HIV-1 replication in vitro. J Virol Methods 1992;38:323.

[12] Huang Z, Olson NA, You W, Haugland RP: A sensitive competitive ELISA for 2,4-dinitrophenol using 3,6-fluorescein diphosphate as a fluorogenic substrate. J Immunol Methods 1992;149:261.

[13] Shalev A, Greenberg AH, McAlpine PJ: Detection of attograms of antigen by a high-sensitivity enzyme-linked immunoabsorbent assay (HS-ELISA) using a fluorogenic substrate. J Immunol Methods 1980;38:125.

[14] Durauer A, Berger E, Schuster M, Wasserbauer E, Himmler G, Loibner Het al. : Peptide arrays for the determination of humoral responses induced by active immunization with a monoclonal antibody against EpCAM. J Immunol Methods 2006;317:114.

[15] Mourtzikou A, Stamouli M, Kroupis C, Christodoulou S, Skondra M, Kastania Aet al. : Evaluation of carcinoembryonic antigen (CEA), epidermal growth factor receptor (EGFR), epithelial cell adhesion molecule EpCAM (GA733-2), and carbohydrate

antigen 19-9 (CA 19-9) levels in colorectal cancer patients and correlation with clinicopathological characteristics. Clin Lab 2012;58:441.

[16] Abe H, Kuroki, M., Imakiire, T., Yamauchi, Y., Yamada, H., Arakawa, F., and Kuroki, M.: Preparation of recombinant MK-1/Ep-CAM and establishment of an ELISA system for determining soluble MK-1/Ep-CAM levels in sera of cancer patients. J Immunol Methods 2002;270:227.

[17] Denzel S, Maetzel D, Mack B, Eggert C, Barr G, Gires O: Initial activation of EpCAM cleavage via cell-to-cell contact. BMC Cancer 2009;9:402.

[18] Gires O: EpCAM a proteolytically cleaved oncogene and an excellent therapeutic target in cancer. Med Sci (Paris). 2009;25:449.

[19] Petsch S, Gires O, Ruttinger D, Denzel S, Lippold S, Baeuerle PAet al. : Concentrations of EpCAM ectodomain as found in sera of cancer patients do not significantly impact redirected lysis and T cell activation by EpCAM/CD3-bispecific BiTE antibody MT110. MAbs 2011;3.

[20] Herlyn D, Somasundaram, R., Zaloudik, J., Jacob, L., Harris, D., Kieny, M.P., Sears, H., and Mastrangelo, M.: Anti-idiotype and recombinant antigen in immunotherapy of colorectal cancer. Cell Biophys 1994;24-25:143.

[21] Herlyn D, Harris, D., Zaloudik, J., Sperlagh, M., Maruyama, H., Jacob, L., Kieny, M.P., Scheck, S., Somasundaram, R., and Hart, E.,: Immunomodulatory activity of monoclonal anti-idiotypic antibody to anti-colorectal carcinoma antibody CO17-1A in animals and patients. J Immunother Emphasis Tumor Immunol 1994;15:303.

[22] Fagerberg J, Frodin, J.E., Wigzell, H., and Mellstedt, H.: Induction of an immune network cascade in cancer patients treated with monoclonal antibodies (ab1). I. May induction of ab1-reactive T cells and anti-anti-idiotypic antibodies (ab3) lead to tumor regression after mAb therapy? Cancer Immunol Immunother 1993;37:264.

[23] Herlyn D, Benden, A., Kane, M., Somasundaram, R., Zaloudik, J., Sperlagh, M., Marks, G., Hart, E., Ralph, C., and Wettendorff, M.: Anti-idiotype cancer vaccines: pre-clinical and clinical studies. In Vivo 1991;5:615.

[24] Herlyn D, Wettendorff, M., and Koprowski, H.: Modulation of cancer patients' immune responses by anti-idiotypic antibodies. Int Rev Immunol 1989;4:347.

[25] Herlyn D, Wettendorff, M., Iliopoulos, D., and Koprowski, H.: Functional mimicry of tumor-associated antigens by antiidiotypic antibodies. Exp Clin Immunogenet 1988;5:165.

[26] Herlyn D, Wettendorff, M., Schmoll, E., Iliopoulos, D., Schedel, I., Dreikhausen, U., Raab, R., Ross, A.H., Jaksche, H., and Scriba, M.: Anti-idiotype immunization of cancer patients: modulation of the immune response. Proc Natl Acad Sci U S A 1987;84:8055.

[27] Herlyn D, Sears H, Iliopoulos D, Lubeck M, Douillard JY, Sindelar Wet al. : Anti-idiotypic antibodies to monoclonal antibody CO17-1A. Hybridoma 1986;5 Suppl 1:S51.

[28] Herlyn D, Ross AH, Koprowski H: Anti-idiotypic antibodies bear the internal image of a human tumor antigen. Science 1986;232:100.

EIDESSTATTLICHE VERSICHERUNG

„Ich, Oliver Schmetzer, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: „Charakterisierung der EpCAM-reaktiven Immunität in Karzinompatienten“. selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.
Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung (siehe „Uniform Requirements for Manuscripts (URM)“ des ICMJE -www.icmje.org) kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) entsprechen den URM (s.o) und werden von mir verantwortet.
Meine Anteile an den ausgewählten Publikationen entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Betreuer/in, angegeben sind. Sämtliche Publikationen, die aus dieser Dissertation hervorgegangen sind und bei denen ich Autor bin, entsprechen den URM (s.o) und werden von mir verantwortet.
Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

_____

Datum Unterschrift

ANTEILSERKLÄRUNG AN DEN ERFOLGTEN PUBLIKATIONEN

Dr. rer. nat. Oliver Schmetzer hatte folgenden Anteil an den vorgelegten Publikationen:

Publikation 1:

Schmetzer O, Moldenhauer G, Riesenberg R, Pires JR, Schlag P, Pezzutto A.

Quality of recombinant protein determines the amount of autoreactivity detected against the tumor-associated epithelial cell adhesion molecule antigen: low frequency of antibodies against the natural protein.

J Immunol. 2005

Anteil: 80 Prozent

Beitrag im Einzelnen: Etablierung und Optimierung des ELISA Systems mit fluorogenen Substrat; Expression, Reinigung und biophysikalische Charakterisierung des nativen EpCAM Antigens (CD und NMR spektroskopie); Sammlung klinischer Proben; Ethikantrag; Messung der Proben; statistische Auswertung und Korrelation; Schreiben des Manuskriptes

Publikation 2:

Schmetzer O, Moldenhauer G, Nicolaou A, Schlag P, Riesenberg R, Pezzutto A.

Detection of circulating tumor-associated antigen depends on the domains recognized by the monoclonal antibodies used: N-terminal trimmed EpCAM-levels are much higher than untrimmed forms.

Immunol Lett. 2012

Anteil: 80 Prozent

Beitrag im Einzelnen: Etablierung des ELISA Systems mit fluorogenen Substrat mit verschieden monoklonalen und polyklonalen Antikörpern; Expression, Reinigung und biophysikalische Charakterisierung der nativen EpCAM Antigens; Charakterisierung anderer mAb-Klone aus Vergleichsstudien (Markierung, Reinigung, FACS-Messung); Sammlung klinischer Proben; Ethikantrag; Messung der Proben; statistische Auswertung; Schreiben des Manuskriptes

Publikation 3:

Riedesel H, Kolbeck B, Schmetzer O, Knapp EW.

Peptide binding at class I major histocompatibility complex scored with linear functions and support vector machines.

Genome Inform. 2004

Anteil: 20 Prozent
Beitrag im Einzelnen: Entwicklung der Fragestellung; Auswertung von MHC-bindenden Peptide aus Literatur und Datenbank als Grundlage für die neue bioinformatischen Modelle; Auswahl von negativ Kontrollpools zur Berechnung; Korrektur des Manuskriptes

Publikation 4:
Salmon D, do Aido-Machado R, Diehl A, Leidert M, Schmetzer O, de A Lima AP, Scharfstein J, Oschkinat H, Pires JR.
Solution structure and backbone dynamics of the Trypanosoma cruzi cysteine protease inhibitor chagasin.
J Mol Biol. 2006
Anteil: 10 Prozent
Beitrag im Einzelnen: Klonierung des Chagasin-Gens in E.Coli, im Expressionsvektor, Expression und Reinigung des Proteins, sowie eines Kontrollproteins; Korrektur des Manuskriptes

Dr. rer. nat. Oliver Schmetzer	Prof. Dr. med. Antonio Pezzutto
Promovend	betreuender Hochschullehrer

Quality of Recombinant Protein Determines the Amount of Autoreactivity Detected against the Tumor-Associated Epithelial Cell Adhesion Molecule Antigen: Low Frequency of Antibodies against the Natural Protein

Oliver Schmetzer,[1]*† Gerhard Moldenhauer,‡ Rainer Riesenberg,§ José Ricardo Pires,‖ Peter Schlag,¶ and Antonio Pezzutto*†

The human epithelial cell adhesion molecule (EpCAM) is expressed on normal epithelial cells and is overexpressed in most carcinomas. EpCAM-targeted immunotherapy has been tried in several clinical studies. High titers of autoantibodies against EpCAM have been reported by different authors. We have generated large amounts of purified protein in S2 *Drosophila* cells (S2-EpCAM) with a purity of >96%. In contrast, the protein produced in baculovirus-dependent systems (baculo-EpCAM) that has been used in previous studies shows a purity of 79%. ^{1}H nuclear magnetic resonance spectrum of S2-EpCAM is typical of folded protein, whereas the baculo-EpCAM sample shows a spectrum corresponding to a partially unfolded protein. Using S2-EpCAM, denatured S2-EpCAM, and baculo-EpCAM, we measured EpCAM Abs of different isotypes in the serum of healthy controls and cancer patients. We found Ab titers against EpCAM in a much lower percentage of sera as published previously, and support the hypothesis that Ab reactivity in some published studies might be due to reactivity against denatured protein, to contaminating proteins in the baculovirus preparations, and to reactivity with BSA. Tetanus toxoid-reactive IgG Abs are present in 1000-fold higher titers compared with EpCAM-reactive Abs. Only IgA Abs were found in higher proportions and in higher concentrations than tetanus toxoid-specific Abs. Our study shows that EpCAM only rarely induces autoantibodies against native protein and emphasizes the importance of using extremely purified Ag preparations when evaluating Abs against tumor-associated Ags. *The Journal of Immunology,* 2005, 174: 942–952.

The epithelial cell adhesion molecule (EpCAM[2]; also termed CO17-1A, GA733-2, KS1-4, EGP, or KSA) has been originally identified as a tumor-associated Ag by the mAb CO17-1A (1). EpCAM is expressed in embryonic tissues and has important functions in the development of pancreas and liver (2–4). It functions as a homotypic cell-cell adhesion molecule and is abundantly expressed in human colon and to a lower extent by nearly all epithelial cells (3–7). In adenocarcinomas, EpCAM can be significantly overexpressed and is a potential target Ag for immunotherapy. Over 90% of all colorectal carcinomas overexpress EpCAM, and the expression is conserved during metastasis (8).

Natural humoral and cellular immune reactivity against EpCAM has been described in cancer patients (9–15), and extensive efforts have been made to develop therapeutic vaccines. Passive immunization with CO17-1A (Panorex) induced survival benefit compared with untreated patients in the clinical setting of minimal residual disease (16, 17). Active immunization has been conducted using anti-Id vaccines (18–20), recombinant protein (13, 21), and viral vectors (22), but has shown limited clinical efficacy so far.

Recombinant proteins contain both CD4 and CD8 epitopes and are potentially useful for vaccination, but their production in large amounts and high purity is not trivial. A frequently used system for the production of rEpCAM rely on the baculovirus infection system, but this yields only low amounts of protein (23). EpCAM production in tobacco plants has been recently described, but it also yields only low amounts of protein and is difficult to establish under good manufacturing practice conditions (24). Therefore, we have expressed EpCAM in S2 *Drosophila* cells (25), trying to establish a good manufacturing practice-suitable procedure. Recombinant protein generated in these cells is usually highly immunogenic due to the presence of mannose-type *N*-glycans, which enhance uptake by APCs and their activation, mediated by mannose receptors (26).

The S2 cells are of embryonic origin and can incorporate large amounts of plasmid DNA (27). They can be cultured in serum-free medium, avoiding contamination with serum proteins.

By optimizing cell culture conditions, transfection rate, and gene expression, we have established a method for obtaining large amounts of highly purified rEpCAM.

Autoantibodies against EpCAM have been demonstrated both in a relatively high percentage of patients with colorectal cancer and in some healthy controls (11, 15). Although the meaning of these autoantibodies is still a matter of debate, it is clear that accurate measurement of Ab titers is crucial for monitoring the immune

*Molecular Immunotherapy, Max Delbrück Centrum for Molecular Medicine, Berlin, †Department of Hematology and Oncology, Charité, Humboldt University, Berlin, ‡German Cancer Research Center, Division of Molecular Immunology (D050), Heidelberg, §Ludwig Maximilians University Muenchen, University Clinic Grosshadern, Department of Urology, Tumor Immunology Group, Munich, ¶Department of Surgical Oncology, Robert Rössle Klinik, Charité, Humboldt University, Berlin, Germany; and ‖Universidade Federal do Rio de Janeiro, Departamento de Bioquimica Medica, Rio de Janeiro, Brazil

Received for publication July 29, 2004. Accepted for publication October 18, 2004.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Address correspondence and reprint requests to Dr. Oliver Schmetzer, Max Delbrück Centrum for Molecular Medicine, Robert-Rössle-Strasse 10, 12135 Berlin, Germany. E-mail address: oliver.schmetzer@gmx.net

[2] Abbreviations used in this paper: EpCAM, epithelial cell adhesion molecule; TT, tetanus toxoid; IEX, ion exchange; NMR, nuclear magnetic resonance; HSA, human serum albumin; DIG, digoxigenin.

0022-1767/05/$02.00

response during immunotherapy. Having planned vaccination with rEpCAM protein, we decided to optimize ELISA measurement of EpCAM autoantibodies. In a first series of blood samples, little or no reactivity against S2-EpCAM was detected, which prompted us to evaluate a large number of sera to clarify discrepancies with published results. To exclude individual variations in the humoral immune response because of the immunosuppressive effects of chemotherapy, we also measured tetanus toxoid (TT)-reactive Abs and compared these titers to the EpCAM-reactive Abs.

Our results indicate that patients with autoantibodies reacting with denatured or misfolded protein may lack Abs against the native form of the Ag. Most autoantibodies are detected by peptide scans or Western blots that would not detect conformational epitopes and may lead to the false assumption that Abs reacting with the native Ag are also present (28).

Materials and Methods

Sera from patients and healthy donors

Sera from 500 patients with carcinomas and from 60 healthy controls were stored in the vapor phase of liquid nitrogen until analyzed. The sera had been collected at Charité University Medicine, Berlin, and Ludwig Maximilians University, Munich, during the past years in the context of different projects. None of the patients received any immunotherapy.

EpCAM-reactive Abs

CO17-1A (Panorex) was obtained from GlaxoSmithKline. HEA125 (29) was produced in a Miniperm bioreactor (Vivascience) and purified by affinity chromatography over protein A-Sepharose CL-4B (Amersham Biosciences). The purity of eluted Ab was assessed by SDS-PAGE under reducing conditions and was >95%. Specificity was proven by indirect immunofluorescence using flow cytometry on living cells and positive staining of Colo205 and SW948 lysates in Western blots (data not shown).

Cell culture

Drosophila S2 cells (Invitrogen Life Technologies) were cultured at 28°C in HyQ SFX Serum (HyClone) supplemented with 10% FBS (FCS), Pen/Strep, gentamicin, pyruvate, and L-glutamine (Biochrom). The serum was only used during selection and expansion of the cells, but not in the protein expression step.

Generation of the rEpCAM from S2 cells

A truncated form of EpCAM was generated by PCR with Pfu-Polymerase (Stratagene) and the following primers: huEpCAM-trunc-5, 5′-AGA TCT ACG GCG ACT TTT GCC GCA GC-3′, and huEpCAM-trunc-3, 5′-TTC GAA TTT TAG ACC CTG CAT TGA GAA TTC-3′, from a cDNA library from human colon. The PCR conditions were 30 cycles of 45 s at 96°C, 45 s at 65°C, and 120 s at 72°C. The PCR product was separated on a 1% agarose gel and excised. The DNA was isolated with QIAEX (Qiagen) and ligated in a PCR Script vector (Stratagene). TOP10 *Escherichia coli* (Invitrogen Life Technologies) were transformed with the ligation product and plated with 100 μg/ml ampicillin. Plasmid minipreps (Qiagen) were analyzed by restriction digestion with *Bgl*II and *Bst*BI (MBI Fermentas) and were sequenced with the following primers: EpCAMs, 5′-agc gag tga gaa cct act gg-3′, and EpCAMas, 5′-acg cgt tgt gat ctc ctt ctg-3′. DNA from a mutation-free clone and from the pMT/BIP/V5-His vector (Invitrogen Life Technologies) was digested with *Bgl*II and *Bst*BI. The 0.75-kB EpCAM insert was ligated in the vector after gel extraction, and the product was used in a transformation of TOP10 *E. coli.* Miniprep DNA was analyzed by digestion with *Bgl*II and *Bst*BI. DNA from one positive clone and from TOP10 *E. coli*, transformed with pCoHygro (Invitrogen Life Technologies), was prepared endotoxin-free with a Maxi-Prep Endofree kit (Qiagen). The pMT-EpCAM vector was sequenced with the EpCAMs and EpCAMas primers and used to transfect S2 cells with Effectene (Qiagen): 2 μg of pMT-EpCAM DNA was mixed with 4 ng of pCoHygro DNA in 200 μl of EC buffer. After addition of 16 μl of Effectene enhancer and incubation for 10 min at room temperature, 100 μl of Effectene reagent was added. Following incubation for 10 min at room temperature, the mix was added to 0.5×10^6 S2 cells in 5 ml of medium. The cells were cultured in a 60-mm dish for 48 h. Selection was started with a concentration of 300 μg/ml hygromycin B (Invitrogen Life Technologies). After 10 days, the concentration was increased to 600 μg/ml hygromycin B and, after a further 10 days, to 1000 μg/ml. The cells were diluted 1/5 two times a week and cultured under these conditions for 12 mo. To induce the expression of the rEpCAM, the cells were washed with serum-free medium and injected in a Celline 1000 bioreactor (Integra Biosciences) at a concentration of 5×10^7/ml. Twenty-five milliliters were injected and 1 liter of serum-free medium was added to the medium compartment of the reactor. Copper sulfate was added to 5 mM after 1 wk. After another week, half of the cell suspension was harvested. This was repeated weekly with exchange of the medium in the medium compartment every 2 wk. The solution was centrifuged twice, and the supernatant was passed through a 0.2-μm filter unit (TPP).

Purification of the rEpCAM

The supernatant from the S2 cells was purified on a Ni-NTA column, packed with 300 ml Ni-NTA-Superflow (Qiagen). After binding, the column was washed with a gradient over 230 ml starting with 100% wash buffer (1.65 M NaCl, 10% glycerol, 0.05% Tween 20, 10 mM Na-phosphate (pH 7.4)) and ending with 96% wash buffer and 20 mM imidazol. Elution was started by applying 200 ml of a gradient, which further increases the imidazol concentration to 500 mM and decreases the amount of wash buffer to 0%. The elution was continued for 100 ml with 500 mM imidazol. For the ion exchange (IEX) chromatography, the protein was dialyzed against 15 mM NaCl and 1 mM Na-phosphate (pH 7.4) in a 6- to 8-kDa dialysis membrane (Spectrum). It was applied to a column with 75 ml of Q Sepharose FF (Amersham Biosciences), and the column was washed with 250 ml of water. Elution was done with a complex gradient, increasing the NaCl concentration over 150 ml to 150 mM. The concentration was further increased with a gradient over 300 ml to 750 mM. This concentration was applied for 200 ml.

Denaturation of the EpCAM protein

EpCAM protein was diluted to a concentration of 0.1 mg/ml in a denaturing buffer (4 M GuHCl, 14.3 mM 2-ME, 15 mM NaCl, 50 mM Tris-Cl (pH 7.4)) and incubated at 37°C for 20 min. The protein solution was dialyzed thereafter against PBS in a 6- to 8-kDa dialysis membrane (Spectrum).

Isolation of EpCAM from the baculovirus expression system

EpCAM produced in the baculovirus expression system was obtained from the MK-1 ELISA kit (Biovendor). According to the specification provided by the company, the standards of the ELISA kit were produced according to the original method described by Strassburg et al. (23).

Nuclear magnetic resonance (NMR) analysis

After dialysis against PBS in a 6- to 8-kDa dialysis membrane (Spectrum), the proteins were concentrated to 10 mg/ml in YM-3 Microcon columns (Millipore). After addition of 50 μl of D_2O to 500 μl of protein solution, the ^1H NMR spectrum was recorded in a Bruker DRX600. The watergate pulse sequence was used to suppress the water signal.

Lectin-binding assay

The lectin-binding assay was performed with the digoxigenin (DIG) Glycan Differentiation kit (Roche) according to the manufacturer's protocol. Briefly, the sample and the control proteins were dotted on a nitrocellulose membrane (Roth). After drying, the membrane was blocked with 2% dry milk powder (Roth) and 0.1% Tween 20 in PBS for 2 h at room temperature. Equal strips were then stained with the different lectin-DIG conjugates, followed by washing with PBS and incubated with the DIG-Fab-enzyme conjugate.

Western blotting

SDS-PAGEs were blotted on nitrocellulose membranes (Schleicher-Schüll) at 1 mA/cm^2 at 4°C for 3 h in a semidry blotting system. Detection of protein on the membrane was done by incubation with amido black solution (0.1% amido black, 25% 2-propanol, and 10% acetic acid) followed by rinsing with water. Blocking was conducted overnight at 4°C with blocking solution (0.5% BSA, 0.2% Tween 20, 5% dry milk powder, 0.1% sodium azide, and 5 mM EDTA in PBS (pH 7.4)). Patients' sera were diluted 1/20 in blocking solution, and 20 ml were used per membrane strip for an incubation overnight at 4°C. After three washes with TPBS (PBS supplemented with 0.1% Tween 20), 50 ml of TPBS with 5 μl each of anti-IgG1-, 2-, 3-, and 4-biotin Abs (BD Pharmingen) were added. The membranes were again incubated overnight and washed three times, and 50 ml of TPBS with 5 μl of avidin-HRP (BD Pharmingen) was added for 2 h at 4°C. Detection was done after four washes with the SuperSignal West Femto Maximum Sensitivity substrate (Pierce) by exposure of 1–5 s to an x-ray film.

ELISAs

Black FluoroNunc MaxiSorb Plates (Nunc) were coated with 250 ng of protein in 100 μl of PBS per well. After incubation overnight at 4°C, the plates were washed once with TPBS, and 400 μl of blocking buffer (2% high purity human serum albumin (HSA; Calbiochem; highest purity), 150 mM NaCl, 0.1% Tween 20, 10 mM Na-phosphate (pH 7.4)) was added to each well. The plates were again incubated overnight and washed once, and patients' sera were added in a geometric dilution in TPBS (1/10 to 1/80). Following overnight incubation, the plates were washed three times, and 100 μl of anti-Ig-biotin Ab (BD Pharmingen) at a concentration of 0.5 μg/ml was added. After 2 h at 4°C, the plates were washed three times, and 100 μl of 0.5 μg/ml streptavidin-AKP (BD Pharmingen) was added. Two hours later, the plates were washed again four times, and 200 μl of substrate solution (0.2 mM 4-methylumbelliferyl phosphate (Sigma-Aldrich), 0.05 M $NaCO_3$, 5 mM $MgCl_2$) was added. The plates were measured after 90 min at room temperature in a Victor II Spectrophotometer (Wallac) with an excitation filter at 365 nm and an emission filter at 450 nm.

The total amounts of Ab were calculated from a titration of purified anti-EpCAM Ab HEA125. The secondary Ab was a biotinylated anti-mouse-IgG1 Ab (BD Pharmingen).

Thirty samples could not be analyzed for IgG3 because BD Pharmingen stopped the production of the necessary Ab.

In an additional experiment, we blocked with BSA (fraction V; protease and nuclease free; Calbiochem) instead of HSA. The samples were considered positive when a fluorescence count >500,000 was reached. This is 3-fold the signal from the negative control (PBS).

Results

Purification of EpCAM protein from S2 cells

EpCAM protein was purified on a Ni-NTA column using an imidazole gradient for elution. The supernatant collected from the bioreactors showed two peaks. The first eluted at a concentration of ~50 mM imidazole and the second at a concentration of 190 mM. Analysis by PAGE showed several proteins with a molecular mass of 40–70 kDa within the first peak and two bands of 32 and 33 kDa in the second elution peak (Fig. 1*A*). The whole fractions of the first peak were discarded; fractions of the second peak were subjected to further purification on a Q Sepharose matrix. The elution resulted in two different protein peaks, whereas some protein did not bind to the matrix. The first peak eluted at 130 mM salt concentration and the second at 380 mM. In the first peak, we

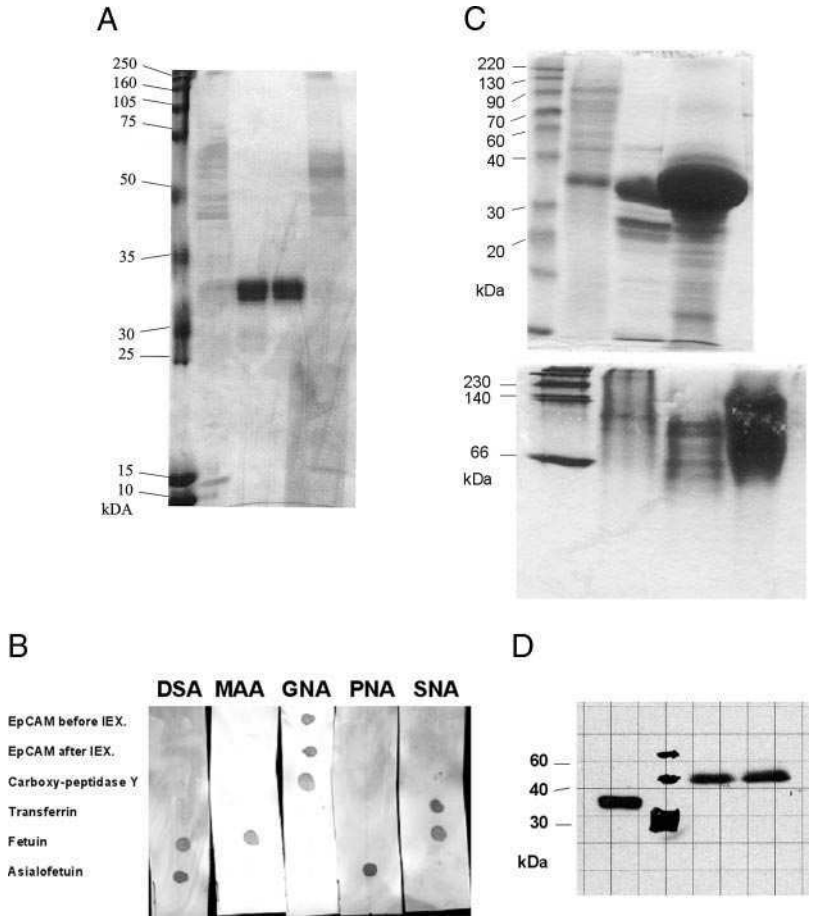

FIGURE 1. *A*, Analysis of rEpCAM protein from S2 cells. PAGE analysis of proteins from peaks eluted during purification. *Lane 1* shows marker proteins. The proteins eluting from the Ni-NTA column at low imidazole concentration are shown in *lane 2*, whereas *lane 3* shows the proteins eluting at higher imidazole concentration. *Lane 4* shows the protein from the IEX column at medium salt concentration. In *lane 5*, the proteins eluting at high salt concentration are shown. *B*, Lectin binding of the EpCAM protein and of control proteins (dot blot). GNA (*Galanthus nivalis* agglutinin) recognizes terminal mannose; SNA (*Sambucus nigra* agglutinin) recognizes sialic acid linked α(2–6) to galactose; MAA (*Maackia amurensis* agglutinin) recognizes sialic acid linked α(2–3) to galactose; PNA (peanut agglutinin) recognizes the core disaccharide galactose β(1–3)*N*-acetylgalactosamine; and DSA (*Datura stramonium* agglutinin) recognizes Galβ-(1–4)*N*-acetylglucosamine (GlcNAc) in complex and hybrid *N*-glycans, in *O*-glycans, and GlcNAc in *O*-glycans. EpCAM appears to contain mannose residues. *C*, Low (10-μg) and high (100-μg) concentrations of purified EpCAM protein were loaded on a PAGE under denaturing (*upper gel*) and native (*lower gel*) conditions. *Lane 1* shows a marker mixture. The removed impurities from IEX chromatography, low and high amounts of purified EpCAM are loaded in *lane 2*, *3*, and *4*, respectively. *D*, Western blot analysis of purified EpCAM with HEA125 as detection Ab. *Lane 1* shows the purified rEpCAM. A protein marker is shown in *lane 2*. *Lanes 3* and *4* show lysates from 293T cells that were infected with two different clones of recombinant adenovirus that carry a full-length *EpCAM* gene.

found again the two 32- and 33-kDa proteins, whereas the second peak yielded a band at ~60 kDa (Fig. 1*A*).

N-Terminal Edman Sequencing of the 32- and 33-kDa protein sample yielded the EpCAM sequence TFAAAQEECVCEN (data not shown). Binding of GNA-lectin in a dot blot showed *N*-glycosylation with mannose-type glycans (Fig. 1*B*).

Multimerization of EpCAM can occur at high protein concentration as shown by native PAGE (Fig. 1*C*) and gel filtration (data not shown). The protein was recognized by HEA125 in a Western blot (Fig. 1*D*) and by CO17-1A in ELISA (data not shown). Serum of animals (rabbits, mice) immunized with the protein stained strongly EpCAM-positive cell lines (Colo205) and EpCAM-cDNA-transfected cell lines (MCA205, B16/F10, and EL4) (data not shown). Spleen cells from mice immunized five times with purified EpCAM protein were used for cell fusion to generate new high-affinity mAbs. These recognize distinct epitopes of the EpCAM protein and can be used for tissue staining, Western blot, and ELISA (our manuscript in preparation).

The total amount of rEpCAM from a single bioreactor was 15–20 mg per harvest.

Comparison of the EpCAM protein purified from S2 cells and from baculovirus-infected cells

Although only the 32- and 33-kDa bands are detected in the eluates of columns loaded with native and denatured EpCAM protein collected from S2 cells supernatants (hereafter called S2-EpCAM), the protein from the commercial kit (hereafter called baculo-EpCAM) contained at least five more proteins with a molecular mass of ~28, 50, 70, 105, and 260 kDa, according to gel staining (Fig. 2*A*). The detection limit of the protein gel is ~0.5 μg of protein as determined by BSA standard dilution (data not shown). Because no bands other than 32–33 kDa were found by loading a total of 12 μg of S2-EpCAM protein, we calculated that the purity of the S2 protein should be >96% (also verified by HPLC analysis, data not shown). The five additional bands in the baculo-EpCAM preparation indicate most likely contaminating proteins with a total mass of at least 2.5 μg as indicated by the gel. The purity of EpCAM in this preparation therefore seems to be no better than 79%. We tried to extract protein from these bands for protein identification by mass spectrometry, but the amounts in the bands were too low. Fig. 2, *B* and *C*, show, respectively, the aliphatic and aromatic/amide regions of the ^{1}H NMR spectrum for S2-EpCAM (lower spectrum), baculo-EpCAM (middle spectrum), and denatured S2-EpCAM (upper spectrum). The ^{1}H NMR spectrum for S2-EpCAM shows comparatively sharp lines and good signal dispersion with several peaks <0.5 ppm and >9.0 ppm. This spectrum is very characteristic of a folded protein with a size of ~30 kDa. The spectrum of baculo-EpCAM otherwise shows very broad lines, suggesting the presence of molecules of higher molecular size, larger proteins or aggregates. In the spectrum of baculo-EpCAM, the dispersion of chemical shifts is also not evident and could indicate the presence of some unfolded protein. The spectrum of the denatured S2-EpCAM (generated by exhaustive reduction of S2-EpCAM) shows the smallest chemical shift dispersion, typical of unfolded proteins. Interestingly, baculo-EpCAM seems to be in an intermediate situation between S2-EpCAM and denatured S2-EpCAM.

Together, these results suggest that the S2-EpCAM consists of a mixture of molecules whereby most are in a native folding state compared with the baculovirus-derived protein, where at least some molecules have spectral features of a denatured protein.

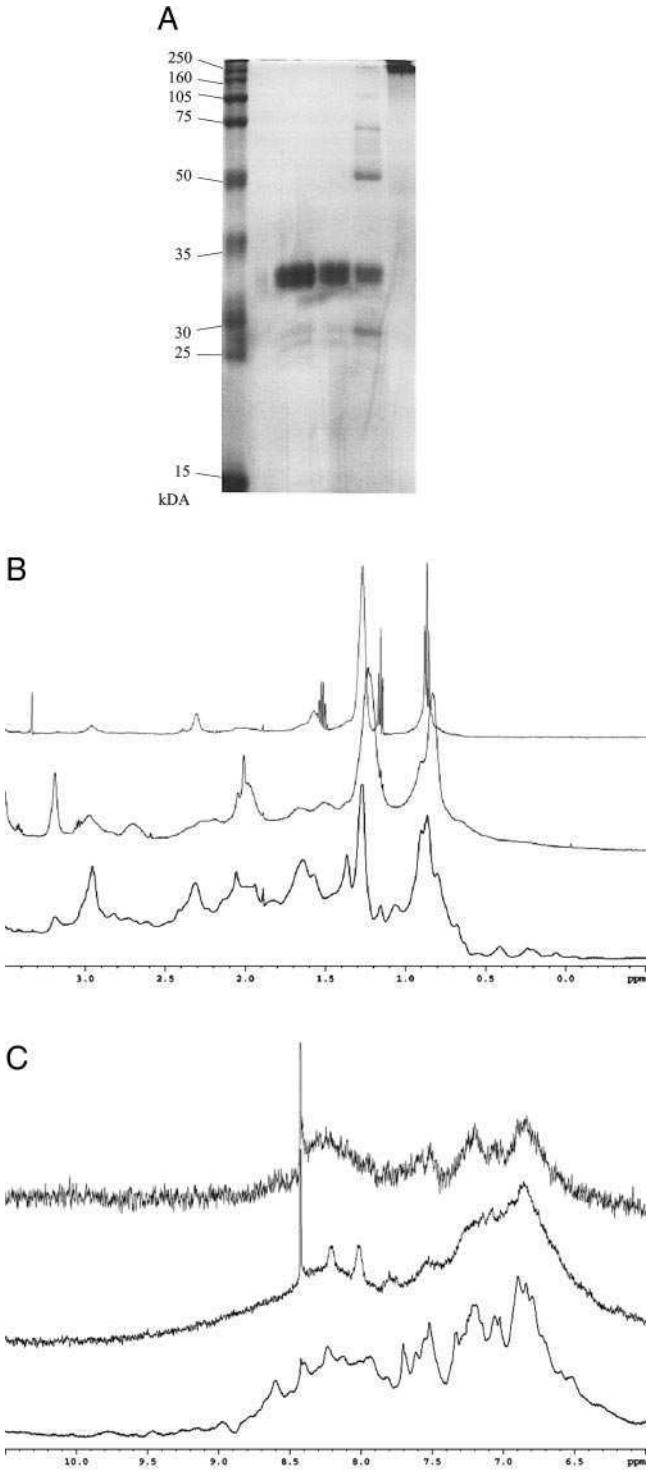

FIGURE 2. *A*, Comparison of EpCAM produced in S2 cells and in baculovirus-infected cells. PAGE analysis of the different test Ags. Marker mixture, native S2-EpCAM, denatured S2-EpCAM, baculo-EpCAM, and TT are shown in *lanes 1*, *3*, *4*, *5*, and *6*, respectively. *Lane 2* is empty. *B*, Aliphatic region of the NMR spectrum from denatured S2-EpCAM (*upper spectrum*), baculo-EpCAM (*middle spectrum*), and native S2-EpCAM (*lower spectrum*). *C*, Aromatic region of the NMR spectrum in the same order as in *B*.

Comparison of autoantibody levels detected with S2-EpCAM and baculo-EpCAM

One hundred fifteen randomly selected sera and five sera that had tested positive in ELISA with S2-EpCAM were further analyzed

for comparative reactivity with S2-EpCAM, denatured S2-EpCAM, baculo-EpCAM, and TT. Sera were considered positive for S2-EpCAM-reactive IgG Abs when the Ab concentration was >80 ng/ml. This amount (8-fold SD of the mean) was never found in 60 sera from healthy donors, as described below.

Abs against baculo-EpCAM were also found in healthy controls (data not shown) as described previously by others (30). As the Ab distribution shows, a threshold for distinguishing positive and negative samples cannot be defined.

All samples were positive for TT-reactive IgG Abs. Forty-five samples showed high amounts of IgG with concentrations exceeding 1 μg/ml. The IgG response consisted mainly of IgG1 and IgG4 Abs; IgG2 responses were often found in lower amounts. Only low levels of IgG3 Abs were detected (Fig. 3*A*). The mean level of IgG Abs against TT was ~1000 times higher compared with S2-EpCAM-reactive Abs. IgM Abs against TT were in the same range as IgM Abs against S2-EpCAM, whereas IgA Abs were more frequently found against EpCAM than against TT, with a mean 3-fold increase.

IgG1, IgG2, and IgG4 Abs against baculo-EpCAM were always present in much higher amounts compared with S2-EpCAM (Fig. 3, *B–G*). There was no correlation between IgG1, IgG2, and IgG4 Abs against S2-EpCAM and baculo-EpCAM. In contrast, similar reactivity was found for IgG3, IgM, and IgA Abs against S2-EpCAM and baculo-EpCAM. When the five selected sera that had tested positive for S2-EpCAM Abs in a preliminary screening were tested again with both reagents, all cases confirmed positive with S2-EpCAM but gave a much higher signal with baculo-EpCAM. No sera tested positive with the S2-EpCAM in the absence of reactivity with baculo-EpCAM.

In 120 tested samples, Abs reacting with denatured S2-EpCAM were detected at higher levels than against native S2-EpCAM with a mean 12-fold increase. Some samples even showed a 100 times higher Ab titer against the denatured preparation (data not shown).

There was no correlation between Ab presence and the age of the patients.

Immune reactivity against EpCAM and contaminating proteins

We further tried to demonstrate specific reactivity against EpCAM in both recombinant protein preparations and against the putative contaminating proteins in the baculo-EpCAM preparation by Western blotting. The staining of the membrane with amido black showed a thick band of EpCAM in the S2-EpCAM preparation, but the background was too high to detect the weak bands in the baculo-EpCAM preparation as expected because of the low sample load (not shown). For detection of Abs, we used five samples that had tested positive against baculo-EpCAM and/or denatured S2-EpCAM in ELISA. We detected immune reactivity against the S2-EpCAM in all samples (Fig. 4*A* and Table I). However, no reactivity against the baculo-EpCAM preparation could be found. The samples contained between 30 and 200 ng of Ab, reactive against denatured S2-EpCAM as determined by ELISA. They also contained 500–700 ng of Abs reactive to the baculo-EpCAM preparation.

We also tested sera from healthy controls for S2-EpCAM reactivity before and after IEX chromatography (Fig. 4*B*). Abs could be found in nearly all samples against the S2-EpCAM preparation without IEX chromatography. These Abs could not be detected after IEX chromatography.

Autoantibody levels detected with S2-EpCAM

In 60 sera from healthy donors, there were only very low levels of S2-EpCAM-reactive IgG Abs (Fig. 5, *A—D*, and Table II). The highest amounts were 8 ng/ml IgG1, 23 ng/ml IgG2, 11 ng/ml IgG3, and 1.8 ng/ml IgG4; however, sera of healthy individuals contained IgM and IgA Abs (Fig. 5, *E* and *F*, and Table II), the highest levels being 172 ng/ml IgM and 84 ng/ml IgA.

Samples were considered positive when the Ab amount exceeded the mean value of the control sera plus 8 SDs. Using this definition, all control sera were negative. Ab amounts higher than 8 ng/ml for IgG1, 27 ng/ml for IgG2, 20 ng/ml for IgG3, 3 ng/ml for IgG4, 371 ng/ml for IgM, and 133 ng/ml for IgA were considered positive according to this definition.

Four of 95 sera from colon cancer patients contained IgG, IgA, or IgM autoantibodies in ELISA using S2-EpCAM (Fig. 5, *A–D*, and Table II). One probe contained 15 ng/ml Ab of IgG1 isotype. One contained IgG2 isotype Abs (160 ng/ml). Abs of IgG4 isotype could be detected in another sample at a level of 41 ng/ml. One patient had IgA Abs: 0.3 μg/ml (Fig. 5*F* and Table II). Among 81 sera from patients with rectal cancer, one had IgG1 (14 ng/ml), two had IgG2 (160 and 221 ng/ml), and three had IgG4 Abs (4.5 and 47 ng/ml) (Fig. 5, *A—D*, and Table II). Three other sera contained IgA Abs (166, 284, and 363 ng/ml) (Fig. 5*F* and Table II). One of the sera was positive for both IgG4 and IgA, whereas all of the others were positive for only one Ab isotype.

Three of 39 sera from patients with gastric cancer had higher Ab amounts compared with the control sera (Fig. 5, *A–D*, Table II). One had IgG1, IgG2, IgG4, and IgA Abs: 79, 121, 22, and 207 ng/ml, respectively. Another patient was positive for IgG4 and IgA (4 and 203 ng/ml, respectively) The third patient had IgG2 Abs only (30 ng/ml). Of 261 sera from patients with breast cancer, six were positive for IgG1 (37, 29, 24, 13, 12, and 10 ng/ml), five were positive for IgG2 (310, 206, 65, 57, and 54 ng/ml), one for IgG3 (23 ng/ml), eight for IgG4 (58, 31, 26, 18, 15, two times 5 and 4 ng/ml), and four for IgA (275, 266, 258, and 172 ng/ml) (Fig. 5 and Table II). Among 32 sera obtained from patients with prostate carcinoma, one tested positive for IgG1 (146 ng/ml) and three for IgA (355, 225, and 158 ng/ml) (Fig. 5 and Table II).

In another experiment, we measured 107 sera from patients with colorectal, stomach, or breast cancer whereby the ELISA plates were blocked with BSA (Table III). All sera had Abs of the IgM isotype. Sixty-one percent of the samples were positive for IgA. IgG was found in 50% of the sera from patients with gastric carcinoma, in 36% of the sera from patients with breast cancer, in 21% of sera from patients with colon cancer, and in 19% of patients with rectal carcinoma.

All samples were measured as duplicates with a SD <5%; all four dilutions showed the expected reduction in the signal strength. All positive sera and 30 randomly selected negative sera were tested again in at least one completely independent experiment with similar results. No correlation with the age of the patients was found in all tested sera. Although clinical data were available only for a minority of the sera, there was no indication of a possible correlation with the disease stage.

Discussion

The presence of autoantibodies against the tumor-associated EpCAM Ag in both cancer patients and normal individuals has been reported widely in the literature. Several immunotherapeutic approaches targeting EpCAM have been developed over the past 20 years. One of the arguments for targeting this Ag despite its wide distribution in the body is the assumption that a certain degree of naturally occurring autoreactivity in cancer patients does not lead to autoimmune organ damage. Autoreactivity against EpCAM is probably representative for a whole range of tumor-associated autoantigens.

In all papers described so far, rEpCAM protein produced in baculovirus systems has been used for detection of autoantibodies.

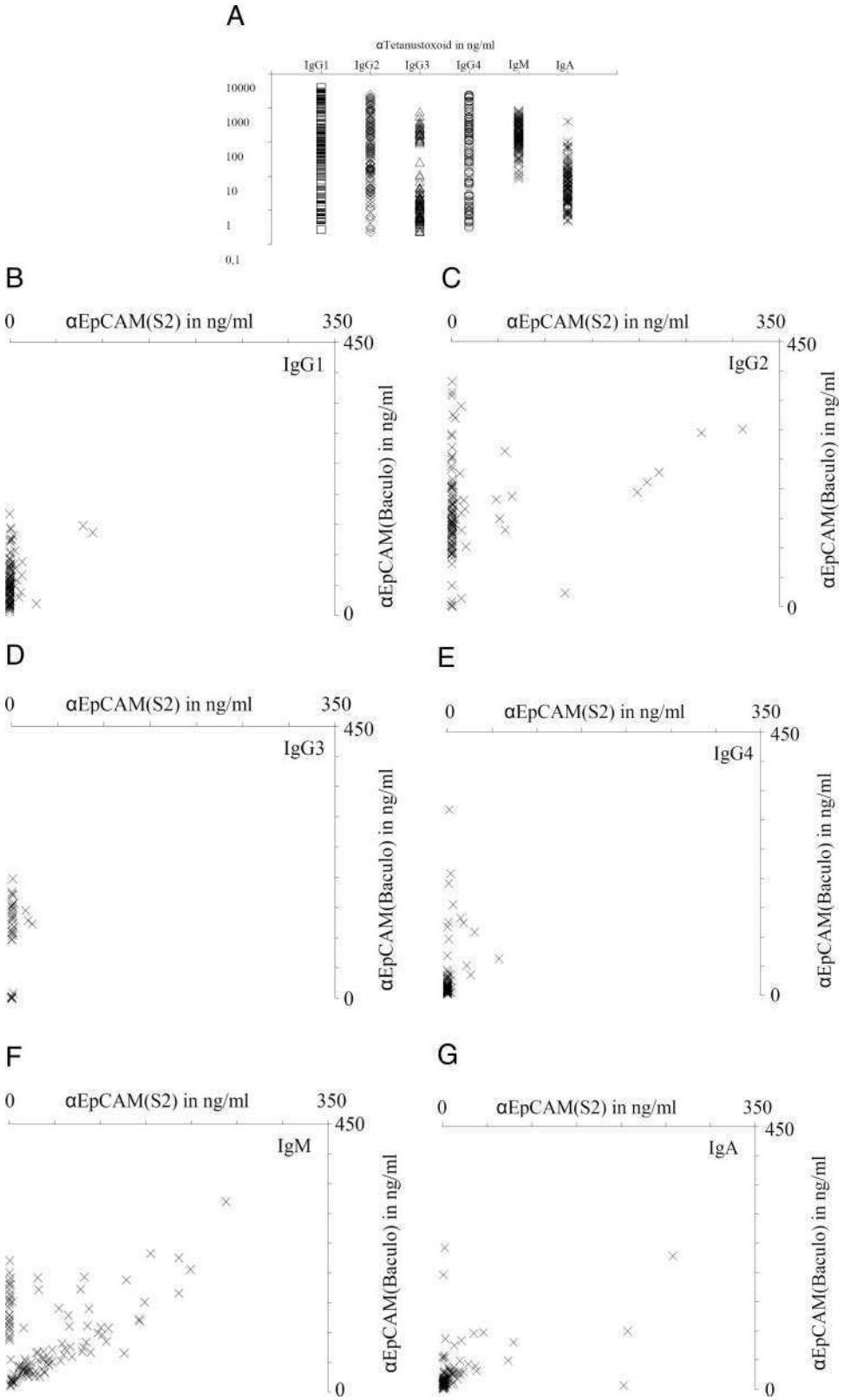

FIGURE 3. *A*, Analysis of the isotypes and amounts of TT-reactive Abs. IgG1 (□), IgG2 (◇), IgG3 (△), IgG4 (○), IgM (×), and IgA (∗) against TT from 120 samples are shown. *B—G*, Comparison of EpCAM-(S2)-reactive and EpCAM-(baculo)-reactive IgG sera. Shown are 115 randomly selected and 5 previously positive tested sera, subtype-specifically analyzed for EpCAM-reactive IgG1, IgG2, IgG3, IgG4, IgM, and IgA autoantibodies.

Even commercial kits have been developed for detection of circulating EpCAM protein in the blood.

Because we envisaged tumor vaccination against EpCAM, highly purified recombinant protein was produced in our lab in S2 *Drosophila* cells. The purity of our protein is significantly higher than the purity of baculovirus EpCAM purified from a commercial preparation (>96 vs ~80%) and was achieved by several optimization steps in the procedure.

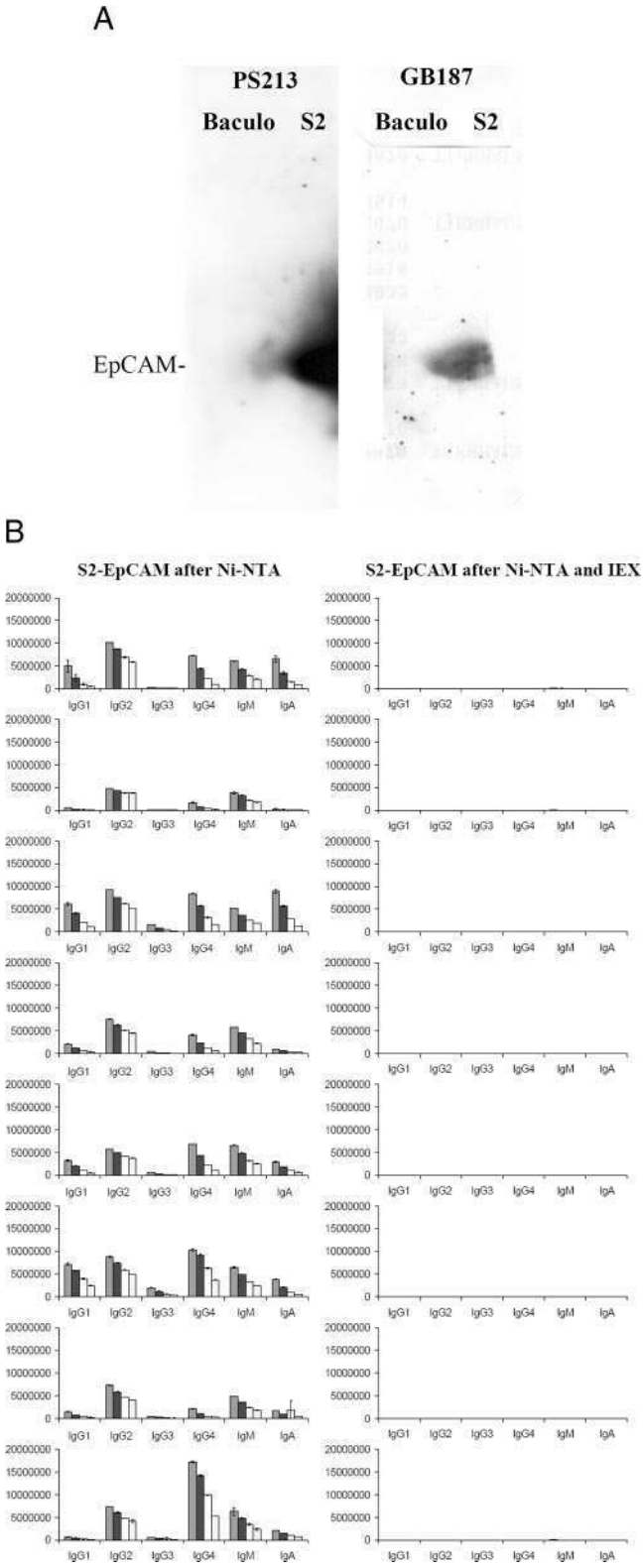

FIGURE 4. *A*, Western blots with patients' sera as primary Ab with either baculo-EpCAM or S2-EpCAM in two lanes on the membrane are shown. *B*, ELISA data from nine healthy controls are shown using EpCAM obtained before and after IEX as test Ag.

Most importantly, our protein appears to be correctly folded according to NMR spectroscopy compared with the baculovirus-produced EpCAM where presence of unfolded or misfolded protein can be assumed.

Analysis by lectin binding indicated that S2-EpCAM is highly mannosylated. N-terminal Edman sequencing verified the purity of the S2-EpCAM protein and showed that removal of the BIP signal peptide was completed in at least 95% of the protein. It also showed that the two 32- and 33-kDa proteins obtained have the same N termini and most likely represent the identical protein, differing only by posttranslational modification, which is in line with our glycosylation data.

Our long selection procedure on hygromycin resulted in an extremely high protein expression by S2 cells; further optimizations of protein expression were achieved by adapting the cells to grow at high concentration in the bioreactors and by optimizing the induction with copper sulfate. This resulted in EpCAM protein concentrations in the supernatant of up to 0.8 mg/ml compared with previous studies where usually 0.5–20 μg/ml recombinant protein was achieved (27, 31, 32).

Comparison of EpCAM purified from S2 cells and baculovirus-derived EpCAM showed major differences. The impurities in the baculovirus preparation become evident only when high amounts of protein are analyzed by PAGE. We detected five different proteins in the baculovirus preparation, which cannot be detected with lower sample loads. These additional bands were either not detected or not described in previous publications. The 70-kDa protein has been discussed as a dimer of EpCAM (23), but the band is visible to the same extent in the original publication, with both reducing and nonreducing conditions. Although it cannot be excluded that dimerization of EpCAM might occur by a mechanism other than disulfide bond formation, it is rather possible that this band represents a contaminating protein. In publications describing the biochemical properties of EpCAM, but not in most publications describing autoantibody reactivity, the baculovirus-derived protein was further purified by at least one further purification step, e.g., by HPLC (33, 34).

Contaminating proteins can produce false-positive results in Ab detection assays. As an example, low titers of Abs can be present in FCS: they can be immobilized and enriched considerably on affinity columns. These Abs or the corresponding immune complexes can elute as contaminating proteins. Furthermore, heterophilic anti-animal Ig Abs (35, 36) can be present in patients' sera and react with BSA or contaminants. The risk of enrichment for contaminating proteins is considerably increased when the concentration of the target protein in the supernatant is low, as is usually the case in baculovirus systems. Although we can purify ~150 mg of protein from 300 ml of supernatant, the protein concentration in the baculovirus system is around 1 mg/L (23).

Moreover, recombinant proteins from baculovirus-infected cells are released upon lysis of the cells by viral infection and not by secretion via the Golgi network, eventually preventing the naturally occurring removal of misfolded protein. This is particularly important for a protein like EpCAM that with its 12 cysteine residues can form many wrong disulfide bridge configurations.

We compared detection of EpCAM-reactive autoantibodies using the baculovirus-derived and the highly purified S2-derived EpCAM protein. Most sera were completely negative if tested with S2-EpCAM, with minimal amounts of Abs being detected in a few sera. In contrast, testing the same sera with baculo-EpCAM showed higher amounts of Abs. This reactivity might be directed against contaminating proteins or misfolded EpCAM protein. To test this, S2-EpCAM was subjected to denaturation and tested in the assay: indeed, 12 times higher amounts of Abs reacted with the denatured EpCAM compared with native S2-EpCAM.

Five sera that tested positive using S2-EpCAM were also positive with baculo-EpCAM as well, indicating that they did recognize EpCAM in both preparations. The recognition of denatured

Table I. *Results from ELISA and Western blotting*[a]

	Patient				
	PS213	KB199	HJ192	GB187	BS205
Anti-EpCAM (S2-nativ) IgG (ng/ml)	3.78	4.32	4.14	30.70	11.91
Anti-EpCAM (S2-denat) IgG (ng/ml)	200.00	66.43	34.83	258.74	55.55
Anti-EpCAM (baculo) IgG (ng/ml)	596.90	725.96	625.31	611.87	616.26
Anti-TT IgG (ng/ml)	702.46	659.47	572.22	396.10	2486.20
Reactivity against S2-EpCAM in WB	+++	++	+	++	++
Reactivity against other proteins in WB	−	−	−	−	−

[a] WB, Western blot.

EpCAM in the S2-EpCAM preparation could also been shown in all tested and ELISA-positive samples by Western blotting, with a good correlation of signal strength and ELISA data. We could not find any reactivity against EpCAM or against the contaminating proteins in the baculo-EpCAM preparation by Western blotting. The Abs reacting against this preparation could recognize nonlinear epitopes, which are destroyed in the SDS gel. Another possibility is the presence of small peptides, large protein aggregates, or nonprotein contaminations that are not separated by the gel. The NMR data suggest the presence of such misfolded large protein aggregates.

The importance of further purification of the test Ag could be shown by performing the ELISA with EpCAM obtained directly after Ni-NTA chromatography and after further purification with IEX chromatography. Only highly purified EpCAM protein showed no reactivity in healthy controls. It is important to mention here that the IEX chromatography removed obviously impurities that could not been seen in SDS-PAGE.

To exclude a falsely low EpCAM reactivity due to selection of patients that have been immunosuppressed by chemotherapy treatment, TT-reactive Abs were also measured in all of the samples. No correlation between EpCAM reactivity and TT reactivity could be found the TT-specific humoral response being ~1000 times higher compared with the EpCAM-specific response. Moreover, TT Abs consisted always of a mixture of several different IgG isotypes, whereas only Abs of a single isotype were found among EpCAM-reactive autoantibodies. These results suggest that the humoral immune response of the patients was not severely compromised by either the tumor itself or the chemotherapy treatment and emphasize the different potency of protective humoral responses against foreign pathogenic Ags with respect to self Ags in patients lacking clinical evidence for autoimmunity.

Among 176 sera from colorectal cancer patients in stage Dukes C or D, 4.5% contained S2-EpCAM-reactive IgGs. One publication describes a frequency of 16% EpCAM-reactive Abs in Dukes C patients and of 32% in Dukes D patients (11). In addition to the false-positive events due to contaminating or misfolded proteins, differences in recruitment of patients and different treatment of control sera could explain the higher frequency in this study. Yet, we do not have a clear explanation for the lack of reactivity of normal sera in this study.

Abs against denatured EpCAM are frequently found: it is possible that due to the presence of proteases and to a certain degree of necrosis and tissue disruption in rapidly growing cancer tissue, denatured protein is generated, which might be more immunogenic than its normal counterpart. Overexpression of the protein, which is typically found in many cancer cells, might also lead to the generation of a higher proportion of misfolded EpCAM, which in turn might be released by damaged cells or undergo increased proteasomal degradation (37). This degradation might also be responsible for the generation of epitopes that could be recognized by T cells, as suggested by our finding that in about one-third of colon cancer patients, T cell reacting against one or more MHC class II binding EpCAM epitopes can be demonstrated.[3] Reactivity against denatured proteins might be at least partially responsible for the reported presence of autoantibodies against many tumor-associated Ags such as cancer-testis Ags. Indeed, in most—if not all—of these studies, denatured protein was used to detect autoantibodies.

This does not necessarily lessen the meaning of these findings, but it may have different implications compared with Abs against correctly folded proteins. Abs against linear epitopes might reflect an epiphenomenon of protein turnover (including formation of misfolded protein and cell death), but are not likely to directly cause cellular damage. This would rather be a possibility with Abs against naturally folded proteins, particularly if expressed on the cell surface. As reviewed recently by Mahler et al. (28) for autoimmunity, disease-causing autoantibodies such as receptor-stimulating Abs against conformational epitopes of the TSH receptor are only found in Graves disease patients and not in normal individuals (38) and anti-glomerular basal membrane Abs that cause Goodpasture disease only react with three-dimensional conformational B epitopes of glomerular basal membrane protein (39). On the contrary, in cancer patients anti-p53 Abs react against a linear peptide that is cryptic in the native p53 and accessible only on the denatured or mutant protein (40).

In a few publications, similar low frequencies of EpCAM autoantibodies as in our study were reported for colorectal cancer patients (18).

Besides reactivity against denatured Ag, all tested sera showed some reactivity to the commonly used BSA. This reactivity is mainly due to IgM Abs, but in many samples, also BSA-reactive IgGs were detected. For this reason we used only highly purified HSA, which resulted in lack of unspecific reactivity. In all studies that measured Abs against EpCAM, BSA was used for blocking, and therefore some degree of BSA reactivity can be assumed. This could also provide an explanation for the high background and the flat profile of the serum dilution curves in some of these studies. Indeed, several publications have shown that BSA-reactive IgG Abs are present in various amounts in 20–100% of sera from different age groups (41–44).

A significant higher amount of EpCAM-reactive IgM was never found in patients with colorectal carcinoma. In healthy controls

[3] O. Schmetzer, M. Schnitzler, G. Richter, W. Zheng, A. Nicolau, H. Koop, B. Hoppe, P. Schlag, F. Sinigaglia, and A. Pezzuttó. Colon cancer patients have circulating T cells that specifically recognize several MHC class II peptides of the tumor-associated EpCAM antigen. *Submitted for publication.*

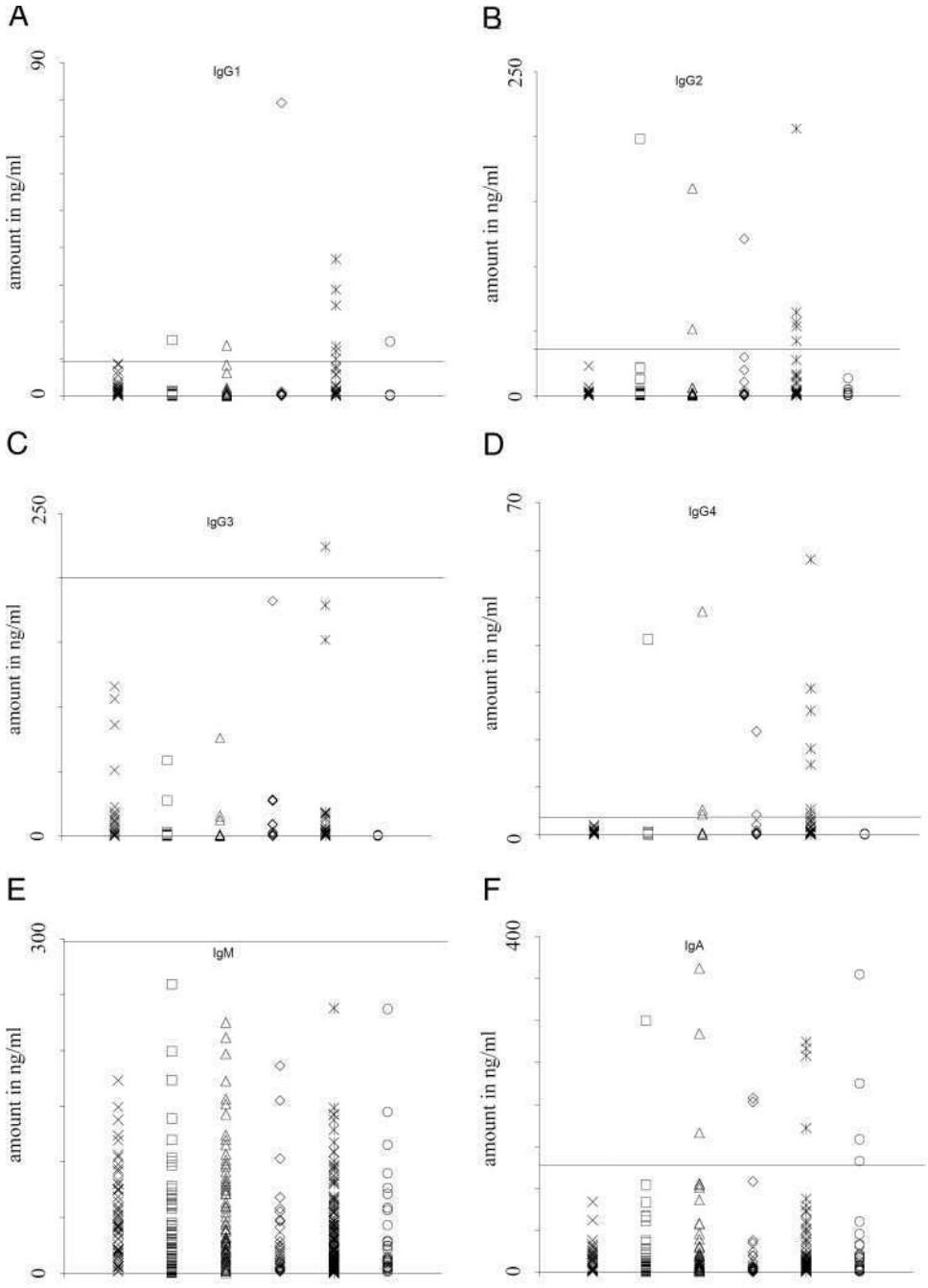

FIGURE 5. Naturally occurring EpCAM-reactive Abs in patients and healthy controls. Shown are 60 sera from healthy control (×), 95 sera from patients with colon carcinoma (□), 81 sera from patients with rectum carcinoma (△), 39 sera from patients with stomach carcinoma (◇), 261 sera from patients with breast cancer (*), and 32 sera from patients with prostate carcinoma (○). The threshold (horizontal line) for positivity has been set at the mean of the sera plus 8 SDs. *A*, EpCAM-reactive IgG1 Abs. *B*, EpCAM-reactive IgG2 Abs. *C*, EpCAM-reactive IgG3 Abs. *D*, EpCAM-reactive IgG4 Abs. *E*, EpCAM-reactive IgM Abs. *F*, EpCAM-reactive IgA Abs.

Table II. *Ab responses (HSA blocking)*[a]

Tumor Type	Number of Samples (*n*)	IgG1 Positive (% (No.))	IgG2 Positive (% (No.))	IgG3 Positive (% (No.))	IgG4 Positive (% (No.))	Total IgG-Positive Samples (% (No.))	Samples Positive for Two or More IgG Isotypes (% (No.))	IgM Positive (% (No.))	IgA Positive (% (No.))
Healthy controls	60	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
Colon CA	95	1.1 (1)	1.1 (1)	0.0 (0)	1.1 (1)	3.2 (3)	0.0 (0)	0.0 (0)	1.1 (1)
Rectum CA	81	1.2 (1)	2.5 (2)	0.0 (0)	3.7 (3)	6.2 (5)	1.2 (1)	0.0 (0)	3.7 (3)
Stomach CA	39	2.6 (1)	5.1 (2)	0.0 (0)	5.1 (2)	7.7 (3)	2.6 (1)	0.0 (0)	5.1 (2)
Mamma CA	261	2.3 (6)	1.9 (5)	0.4 (1)	3.1 (8)	6.9 (18)	0.8 (2)	0.0 (0)	1.5 (4)
Prostate CA	32	3.1 (1)	0.0 (0)	0.0 (0)	0.0 (0)	3.1 (1)	0.0 (0)	0.0 (0)	9.4 (3)

[a] CA, Cancer.

Table III. *Ab responses (BSA blocking)*[a]

Tumor Type	Number of Samples *n*	IgG1 Positive (% (No.))	IgG2 Positive (% (No.))	IgG3 Positive (% (No.))	IgG4 Positive (% (No.))	Total IgG-Positive Samples (% (No.))	Samples Positive for Two or More IgG-Isotypes (% (No.))	IgM Positive (% (No.))	IgA Positive (% (No.))
Colon-CA	33	6.1 (2)	15.2 (5)	0.0 (0)	15.2 (5)	21.2 (7)	9.1 (3)	100.0 (33)	48.5 (16)
Rectum-CA	21	0.0 (0)	19.0 (4)	0.0 (0)	0.0 (0)	19.0 (4)	0.0 (0)	100.0 (21)	66.7 (14)
Stomach-CA	8	0.0 (0)	50.0 (4)	0.0 (0)	12.5 (1)	50.0 (4)	12.5 (1)	100.0 (8)	75.0 (6)
Mamma-CA	45	4.4 (2)	33.3 (15)	6.7 (3)	6.7 (3)	35.6 (16)	6.7 (3)	100.0 (45)	64.4 (29)

[a] CA, Cancer.

also similar high amounts of IgM were detectable, which may correspond to autoantibodies of low specificity. Such autoantibodies have been described widely (45).

EpCAM-reactive IgA autoantibodies have never been described previously. We found approximately three times higher amounts of IgA against EpCAM compared with TT. This could be due to the different localization of the Ag. Although TT-reactive Abs are often induced by vaccine injection in muscle tissue, EpCAM is abundantly expressed in the gut, where high concentrations of TGF-β are needed to support the growth of epithelial stem cells. Because TGF-β facilitates IgA class switch (43, 44), B lymphocytes resident in the gut may be induced to preferentially secrete IgA by the local TGF-β-rich environment.

EpCAM-reactive Abs of the IgG3 subclass were present in minimal amounts in only one of 120 samples. This low frequency could be explained by the higher complement-activating capacity of this Ig isotype (46, 47): one could assume that the class switch to IgG3 is more tightly regulated compared with other isotypes to prevent autoimmune tissue damage.

We could find very low levels of EpCAM-reactive Abs in healthy donors. This is consistent with recent findings (30). Because of the high sensitivity of our test, some normal individuals appeared to have EpCAM autoantibodies in low titers if we used 3 SDs as a threshold. We felt that using 8 SDs provided a safe boundary for keeping most of healthy individuals negative and most of the patients who appear to have circulating autoantibodies positive. Arbitrary thresholds have to be set for many autoantibodies such as rheumatoid factors or cold agglutinins, which are present in many normal individuals in low concentrations.

Because of the high expression of EpCAM in several fetal organs, one would expect the existence of central tolerance for this Ag (2, 4, 7). Also, the high expression of EpCAM on nearly all epithelial cells should render T cells that may have escaped the thymic selection anergic and induce a state of peripheral tolerance. Moreover, oral tolerance may be induced by the long-term shedding of dead epithelial cells in the gut. In the thymus, negative selection of CD8 cells is more accurate than for CD4 cells. However, $CD4^+$ T cell responses against EpCAM have been demonstrated (48). We could also define specific MHC-II binding EpCAM epitopes that appear to be recognized by a small percentage of cancer patients.[3] The presence of IgG Abs against EpCAM in some patients indeed fits well to CD4-mediated Ig-class switching.

A few clinical trials have been conducted with the aim of inducing EpCAM-specific immune responses using viral vectors or anti-idiotypic Abs (22, 49–52). There was none or only a minimal benefit to the patient in studies with anti-idiotypic vaccines, although the vaccine induced strong humoral and cellular immune response against the injected Ab. Autoimmunity was never observed.

Our results indicate that Ab reactivity against tumor Ags as claimed by many publications must be cautiously interpreted and clearly depend on the purity of the Ag used for Ab detection. The immune response that, indeed, some cancer patients appear to mount against EpCAM seems to be weak, and the effectivity of strategies aiming at increasing this response might be severely limited by a state of immune tolerance. Breaking tolerance against this tumor-associated Ag might indeed provide antitumor activity but at the price of autoimmunity.

Acknowledgments

We thank T. Blankenstein (Max Delbrück Centrum, Berlin, Germany) and W. Zimmermann (Ludwig Maximilians University, Munich, Germany) for helpful discussion. Prof. H. Oschkinat provided help with NMR spectrometry evaluation.

References

1. Herlyn, M., Z. Steplewski, D. Herlyn, and H. Koprowski. 1979. Colorectal carcinoma-specific antigen: detection by means of monoclonal antibodies. *Proc. Natl. Acad. Sci. USA 76:1438*.
2. Cirulli, V., C. Ricordi, and A. Hayek. 1995. E-cadherin, NCAM, and EpCAM expression in human fetal pancreata. *Transplant. Proc. 27:3335*.
3. Cirulli, V., L. Crisa, G. M. Beattie, M. I. Mally, A. D. Lopez, A. Fannon, A. Ptasznik, L. Inverardi, C. Ricordi, T. Deerinck, et al. 1998. KSA antigen Ep-CAM mediates cell-cell adhesion of pancreatic epithelial cells: morphoregulatory roles in pancreatic islet development. *J. Cell Biol. 140:1519*.
4. de Boer, C. J., J. H. van Krieken, C. M. Janssen-van Rhijn, and S. V. Litvinov. 1999. Expression of Ep-CAM in normal, regenerating, metaplastic, and neoplastic liver. *J. Pathol. 188:201*.
5. Momburg, F., G. Moldenhauer, G. J. Hammerling, and P. Moller. 1987. Immunohistochemical study of the expression of a M_r 34000 human epithelium-specific surface glycoprotein in normal and malignant tissue. *Cancer Res. 47:2883*.
6. Litvinov, S. V., H. A. Bakker, M. M. Gourevitch, M. P. Velders, and S. O. Warnaar. 1994. Evidence for a role of the epithelial glycoprotein 40 (Ep-CAM) in epithelial cell-cell adhesion. *Cell Adhes. Commun. 2:417*.
7. Balzar, M., M. J. Winter, C. J. de Boer, and S. V. Litvinov. 1999. The biology of the 17-1A antigen (Ep-CAM). *J. Mol. Med. 77:699*.
8. Sheyte, J., B. Christensson, C. Rubio, M. Rodensjo, P. Biberfeld, and H. Mellstedt. 1989. The tumor-associated antigens BR55-2, GA73-3 and GICA 19-9 in normal and corresponding neoplastic human tissues, especially gastrointestinal tissues. *Anticancer Res. 9:395*.
9. Herlyn, D., R. Somasundaram, J. Zaloudik, L. Jacob, D. Harris, M. P. Kieny, H. Sears, and M. Mastrangelo. 1994. Anti-idiotype and recombinant antigen in immunotherapy of colorectal cancer. *Cell Biophys. 24–25:143*.
10. Ras, E., S. H. van der Burg, S. T. Zegveld, R. M. Brandt, P. J. Kuppen, R. Offringa, S. O. Warnarr, C. J. van de Velde, and C. J. Melief. 1997. Identification of potential HLA-A*0201 restricted CTL epitopes derived from the epithelial cell adhesion molecule (Ep-CAM) and the carcinoembryonic antigen (CEA). *Hum. Immunol. 53:81*.
11. Mosolits, S., U. Harmenberg, U. Ruden, L. Ohman, B. Nilsson, B. Wahren, J. Fagerberg, and H. Mellstedt. 1999. Autoantibodies against the tumour-associated antigen GA733-2 in patients with colorectal carcinoma. *Cancer Immunol. Immunother. 47:315*.
12. Nagorsen, D., U. Keilholz, L. Rivoltini, A. Schmittel, A. Letsch, A. M. Asemissen, G. Berger, H. J. Buhr, E. Thiel, and C. Scheibenbogen. 2000. Natural T-cell response against MHC class I epitopes of epithelial cell adhesion molecule, her-2/neu, and carcinoembryonic antigen in patients with colorectal cancer. *Cancer Res. 60:4850*.
13. Staib, L., B. Birebent, R. Somasundaram, E. Purev, H. Braumuller, C. Leeser, N. Kuttner, W. Li, D. Zhu, J. Diao, et al. 2001. Immunogenicity of recombinant GA733-2E antigen (CO17-1A, EGP, KS1-4, KSA, Ep-CAM) in gastro-intestinal carcinoma patients. *Int. J. Cancer 92:79*.
14. Trojan, A., M. Witzens, J. L. Schultze, R. H. Vonderheide, S. Harig, A. M. Krackhardt, R. A. Stahel, and J. G. Gribben. 2001. Generation of cytotoxic T lymphocytes against native and altered peptides of human leukocyte antigen-A*0201 restricted epitopes from the human epithelial cell adhesion molecule. *Cancer Res. 61:4761*.

15. Mosolits, S., M. Steinitz, U. Harmenberg, U. Ruden, E. Eriksson, H. Mellstedt, and J. Fagerberg. 2002. Immunogenic regions of the GA733-2 tumor-associated antigen recognised by autoantibodies of patients with colorectal carcinoma. *Cancer Immunol. Immunother. 51:209*.
16. Riethmuller, G., E. Schneider-Gadicke, G. Schlimok, W. Schmiegel, R. Raab, K. Hoffken, R. Gruber, H. Pichlmaier, H. Hirche, and R. Pichlmayr. 1994. Randomised trial of monoclonal antibody for adjuvant therapy of resected Dukes' C colorectal carcinoma: German Cancer Aid 17-1A Study Group. *Lancet 343:1177*.
17. Riethmuller, G., E. Holz, G. Schlimok, W. Schmiegel, R. Raab, K. Hoffken, R. Gruber, I. Funke, H. Pichlmaier, H. Hirche, et al. 1998. Monoclonal antibody therapy for resected Dukes' C colorectal cancer: seven-year outcome of a multicenter randomized trial. *J. Clin. Oncol. 16:1788*.
18. Herlyn, D., R. Somasundaram, J. Zaloudik, W. Li, L. Jacob, D. Harris, M. P. Kieny, R. Ricciardi, E. Gonczol, and H. Sears. 1995. Cloned antigens and antiidiotypes. *Hybridoma 14:159*.
19. Mellstedt, H., J. Fagerberg, J. E. Frodin, A. L. Hjelm-Skog, M. Liljefors, K. Markovic, S. Mosolits, and P. Ragnhammar. 2000. Ga733/EpCAM as a target for passive and active specific immunotherapy in patients with colorectal carcinoma. *Ann. NY Acad. Sci. 910:254*.
20. Somasundaram, R., J. Zaloudik, L. Jacob, A. Benden, M. Sperlagh, E. Hart, G. Marks, M. Kane, M. Mastrangelo, and D. Herlyn. 1995. Induction of antigen-specific T and B cell immunity in colon carcinoma patients by anti-idiotypic antibody. *J. Immunol. 155:3253*.
21. Basak, S., S. Eck, R. Gutzmer, A. J. Smith, B. Birebent, E. Purev, L. Staib, R. Somasundaram, J. Zaloudik, W. Li, et al. 2000. Colorectal cancer vaccines: antiidiotypic antibody, recombinant protein, and viral vector. *Ann. NY Acad. Sci. 910:237*.
22. Ullenhag, G. J., J. E. Frödin, S. Mosolits, S. C. Bonnet, P. Moingeon, H. Mellstedt, and H. Rabbani. 2003. Immunization of colorectal carcinoma patients with a recombinant canarypox virus expressing the tumor antigen Ep-CAM/KSA (ALVAC-KSA) and granulocyte macrophage colony-stimulating fator induced a tumor-specific cellular immune response. *Clin. Cancer Res. 9:2447*.
23. Strassburg, C. P., Y. Kasai, B. A. Seng, P. Miniou, J. Zaloudik, D. Herlyn, H. Koprowski, and A. J. Linnenbach. 1992. Baculovirus recombinant expressing a secreted form of a transmembrane carcinoma-associated antigen. *Cancer Res. 52:815*.
24. Verch, T., D. C. Hooper, A. Kiyatkin, Z. Steplewski, and H. Korprowski. 2004. Immunization with a plant-produced colorectal antigen. *Cancer Immunol. Immunother. 53:92*.
25. Schneider, I. 1972. Cell lines derived from late embryonic stages of *Drosophila melanogaster*. *J. Embryol. Exp. Morphol. 27:353*.
26. Apostolopoulos, V., and I. F. C. McKenzie. 2001. Role of the mannose receptor in the immune response. *Curr. Mol. Med. 1:469*.
27. Benting, J., S. Lecat, D. Zacchetti, and K. Simons. 2000. Protein expression in *Drosophila* cells. *Anal. Biochem. 278:59*.
28. Mahler, M., M. Bluthner, and K. M. Pollard. 2003. Advances in B-cell epitope analysis of autoantigens. *Clin. Immunol. 107:65*.
29. Moldenhauer, G., F. Momburg, P. Möller, R. Schwartz, and G. J. Hämmerling. 1987. Epithelium specific surface glycoprotein of M_r 34,000 is a widely distributed human carcinoma marker. *Br. J. Cancer 56:712*.
30. Kirman, I., D. Jenkins, R. Fowler, and R. L. Whelan. 2003. Naturally occurring antibodies to epithelial cell adhesion molecule (EpCAM). *Dig. Dis. Sci. 48:2306*.
31. Nilsen, S. L., and F. J. Castellino. 1999. Expression of human plasminogen in *Drosophila* Schneider S2 cells. *Protein Expression Purif. 16:136*.
32. Wang, W.-C., K. Zinn, and P. J. Bjorkmanns. 1993. Expression and structural studies of fasciclin I, and insect adhesion molecule. *J. Biol. Chem. 268:1448*.
33. Trebak, M., G. E. Begg, J. M. Chong, E. V. Kanazireva, D. Herlyn, and D. W. Speicher. 2001. Oligomeric state of the colon carcinoma-associated glycoprotein GA733-2 (Ep-CAM/EGP40) and its role in GA733-mediated homotypic cell-cell adhesion. *J. Biol. Chem. 276:2299*.
34. Chong, J. M., and D. W. Speicher. 2001. Determination of disulfide bond assignments and N-glycosylation sites of the human gastrointestinal antigen GA733-2. *J. Biol. Chem. 276:5804*.
35. Andersson, M., J. Rönnmark, I. Areström, P. A. Nygren, and N. Ahlborg. 2003. Inclusion of non-immunoglobulin binding protein in two-site ELISA for quantification of human serum proteins without interference by heterophilic serum antibodies. *J. Immunol. Methods 283:225*.
36. Henning, C., L. Rink, U. Fagin, W. J. Jabs, and H. Kirchner. 2000. The influence of naturally occurring heterophilic anti-immunoglobulin antibodies on direct measurement of serum proteins using sandwich ELISAs. *J. Immunol. Methods 235:71*.
37. Wickner, S., M. R. Maurizi, and S. Gottesman. 1999. Posttranslational quality control: folding, refolding, and degrading proteins. *Science 286:1888*.
38. Atger, M., M. Misrahi, J. Young, A. Jolivet, J. Orgiazzi, G. Schaison, and E. Milgrom. 1999. Autoantibodies interacting with purified native thyrotropin receptor. *Eur. J. Biochem. 265:1022*.
39. Wu, J., J. Arends, J. Borillo, C. Zhou, J. Merszei, J. McMahon, and Y. H. Lou. 2004. A self T cell epitope induces autoantibody response: mechanism for production of antibodies to diverse glomerular basement membrane antigens. *J. Immunol. 172:4567*.
40. Lubin, R., B. Schlichtholz, D. Bengoufa, G. Zalcman, J. Tredaniel, A. Hirsch, C. C. de Fromentel, C. Preudhomme, P. Fenaux, G. Fournier, et al. 1993. Analysis of p53 antibodies in patients with various cancers define B-cell epitopes of human p53: distribution on primary structure and exposure on protein surface. *Cancer Res. 53:5872*.
41. Hilger, C., F. Grigioni, C. De Beauford, G. Michel, J. Freilinger, and F. Henges. 2001. Differential binding of IgG and IgA antibodies to antigenic determinants of bovine serum albumin. *Clin. Exp. Immunol. 123:387*.
42. Pardini, V. C., W. Miranda, S. R. Ferreira, G. Vehlo, and E. M. Russo. 1996. Antibodies to bovine serum albumin in Brazilian children and young adults with IDDM. *Diabetes Care 2:126*.
43. Schonheyder, H., and P. Andersen. 1984. Effects of bovine serum albumin on antibody determination by the enzyme-linked immunosorbent assay. *J. Immunol. Methods 72:251*.
44. Levy-Marchal, C., J. Karjalainen, F. Dubois, W. Karges, P. Czernichow, and H. M. Dosch. 1995. Antibodies against bovine serum albumin and other diabetes markers in French children. *Diabetes Care 18:1089*.
45. Schattner, A. 1987. The origin of autoantibodies. *Immunol. Lett. 14:143*.
46. Brekke, O. H., T. E. Michaelsen, and I. Sandlie. 1995. The structural requirements for complement activation by IgG: does it hinge on the hinge? *Immunol. Today 16:85*.
47. Feinstein, A., N. Richardson, and M. J. Taussig. 1986. Immunoglobulin fexibility in complement activation. *Immunol. Today 7:169*.
48. Herlyn, D., M. Wettendorff, E. Schmoll, D. Iliopoulos, I. Schedel, U. Dreikhausen, R. Raab, A. H. Ross, H. Jaksche, and M. Scriba. 1987. Anti-idiotype immunization of cancer patients: modulation of the immune response. *Proc. Natl. Acad. Sci. USA 84:8055*.
49. Birebent, B., R. Somasundaram, E. Purev, W. Li, E. Mitchell, D. Hoey, E. Bloom, M. Mastrangelo, H. Maguire, D. T. Harris, et al. 2001. Anti-idiotypic antibody and recombinant antigen vaccines in colorectal cancer patients. *Crit. Rev. Oncol. Hematol. 39:107*.
50. Herlyn, D., M. Wettendorff, D. Iliopoulos, E. Schmoll, I. Schedel, and H. Koprowski. 1989. Modulation of cancer patients' immune responses by administration of anti-idiotypic antibodies. *Viral Immunol. 2:271*.
51. Tempero, M. A., E. Uchida, D. Herlyn, and Z. Steplewski. 1986. Monoclonal antibody CO17-1A and leukapheresis in immunotherapy of pancreatic cancer. *Hybridoma 5(Suppl. 1):S133*.
52. Gruber, R., L. J. M. van Haarlem, S. O. Warnaar, E. Holz, and G. Riethmüller. 2000. The human antimouse immunglobulin reponse and the anti-idiotypic network have no influence on clinical outcome in patients with minimal residual colorectal cancer treated with monoclonal antibody CO17-1A. *Cancer Res. 60:1921*.

Immunology Letters 143 (2012) 184–192

Contents lists available at SciVerse ScienceDirect

Immunology Letters

journal homepage: www.elsevier.com/locate/immlet

Detection of circulating tumor-associated antigen depends on the domains recognized by the monoclonal antibodies used: N-terminal trimmed EpCAM-levels are much higher than untrimmed forms

Oliver Schmetzer[a,b,*], Gerhard Moldenhauer[e], Annett Nicolaou[b], Peter Schlag[c], Rainer Riesenberg[d], Antonio Pezzutto[a,b]

[a] Molecular Immunotherapy, Max-Delbrück-Centrum for Molecular Medicine, Berlin, Germany
[b] Department of Hematology and Oncology, Charité, Humboldt University, Berlin, Germany
[c] Department of Surgical Oncology, Robert-Rössle-Klinik, Charité, Humboldt University, Berlin, Germany
[d] Ludwig-Maximilians-University Muenchen, University Clinic Grosshadern, Department of Urology, Tumor Immunology Group, Münich, Germany
[e] German Cancer Research Center, Translational Immunology Unit (D015), Heidelberg, Germany

ARTICLE INFO

Article history:
Received 13 December 2011
Received in revised form 29 January 2012
Accepted 14 February 2012
Available online 22 February 2012

Keywords:
EpCAM
Colon cancer
Autoantibodies
Enzyme linked immunosorbent assay
Humoral tolerance
Tumor marker

ABSTRACT

The measurement of tumor-associated proteins is of high diagnostic value in the follow-up of cancer patients. Most tests ignore that various forms of the protein can exist; especially in epithelial cancers and the soluble receptors they produce. We choose EpCAM as model-antigen to analyze whether tests recognizing different domains of the protein give different results in patients' sera. EpCAM-reactive autoantibodies are present in the sera of patients with colorectal carcinoma, however little is known about the existence and possible relevance of circulating soluble EpCAM protein. Most monoclonal EpCAM-antibodies recognize the first EGF-like repeat and fail to detect N-terminal trimmed protein.

We developed a novel ELISA to determine the concentration of serum EpCAM with mAbs recognizing the second EGF-like repeat. In 59 healthy controls, EpCAM concentrations ranged from 232 to 8893 ng/ml (mean 1525 ng/ml). Levels of EpCAM in 412 patients with adenocarcinoma were somewhat higher with concentrations ranging from 176 to 36,259 ng/ml (mean 1971 ng/ml). In direct comparison, the untrimmed protein specific ELISA detected lower levels and frequencies as compared to the EGFII-specific ELISA.

Only sera with less than 1 μg/ml circulating EGFII-EpCAM (66% of the sera) contained EpCAM-specific IgG antibodies. The absence of IgG antibodies in the sera with more than 1 μg/ml circulating EpCAM was not due to immune complex formation. Anti-EpCAM IgA and IgM antibodies did not show such a correlation.

It will be important to assess whether the presence of high levels of circulating EGFII-EpCAM is associated with side effects in patients given immunotherapy.

1. Introduction

The epithelial cell adhesion molecule (EpCAM, also known as CD326, KSA, EGP, KS1-4, CO17-1A, GA733-2 or MK-1) was originally identified by the monoclonal antibody (mAb) CO17-1A, which has been used with some success in clinical trials [1–4]. EpCAM is normally expressed in many epithelial tissues throughout the body; elevated expression has been found in carcinomas of the colon, rectum, stomach, breast, prostate, hepatocellular, biliary tract and pancreas [5,6]. High EpCAM expression has been associated with a poorer prognosis in a wide variety of different carcinomas [7–9]. EpCAM is an important homophilic adhesion molecule; its second EGF-like domain is responsible for homodimerization [10,11]. For activation of its intracellular signaling cascade, EpCAM must be cleaved, thereby releasing its extracellular domain called EpEX. Moreover, a splice variant called EpCAM-006 encodes a soluble form of the protein without its transmembrane region. Soluble forms of EpCAM can be detected in human plasma [12–16], which are a mix of shed extracellular EpCAM and the soluble protein encoded by the splice variant. However, relatively little is known about the significance of EpCAM levels in the blood [16–18].

Different concentrations for soluble forms of carcinoma-associated proteins in human serum have been often described,

* Corresponding author at: Charité University Medicine Berlin, Clinic for Hematology and Oncology, Augustenburger Platz 1, 13353 Berlin, Germany.
Tel.: +49 30 9406 2695; fax: +49 30 9406 2698.
E-mail address: oliver.schmetzer@gmx.net (O. Schmetzer).

0165-2478/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.imlet.2012.02.004

explanations for discrepant results mainly focused on specificity and sensitivity of the reagents and not on different soluble forms of the receptors. Only for some antigens the differences have been analyzed in greater detail. That is, it has been shown that the homodimers VEGF121 and VEGF110 cannot be recognized by one ELISA, while the full-length extracellular VEGF165 part can be detected [19]. One study compared different ELISAs for the soluble VEGF-isoforms and found different concentrations depending on the epitopes recognized by the ELISA [20]. Importantly, the different VEGF isoforms have been shown to have different clinical predictive value [21].

Especially in quantification of soluble antigen, the choice of the ELISA-mAbs has been demonstrated to be crucial. Detected Osteopontin levels varied more than a log-scale in sera from breast cancer patients depending on the ELISA used and on which fragment of the protein is detected [22,23].

Depending on the used capture Ab, different concentrations of HLA-G were determined after direct comparison of different ELISAs [24]. Not fragmentation of HLA-G was the reason for this finding, but complex-formation with other proteins.

Because most available EpCAM-antibodies recognize the first epidermal growth factor (EGF)-like domain, we decided to develop a fluorescence-based ELISA targeting the second EGF-like domain to determine concentrations of EpCAM in the sera of cancer patients and healthy controls. In direct comparison much higher levels were found with the EGFII-targeted mAbs as compared to the untrimmed protein-specific ELISA. We also evaluated the relationship between the levels of soluble EpCAM and the presence of serum autoantibodies. We found an inverse correlation between autoantibody and serum EpCAM levels, which is not due to formation of immune complexes.

2. Material and methods

2.1. Sera from patients and healthy donors

588 serum samples were obtained from donors after informed consent as approved by the ethics committee of the Charité Medical School, Berlin in congruence with the 1964 Declaration of Helsinki. 120 samples from healthy donors were derived from the Ludwig-Maximilians-University Munich. In total, sera from 548 patients with adenocarcinoma (blood was drawn after surgery or in-between the first cycles of chemotherapy), from 59 healthy controls and from 37 patients with cardiovascular diseases (with no clinical evidence for cancer) were analyzed to determine the concentration of circulating EpCAM-protein as described below. 401 of the same serum samples had been analyzed for the presence of anti-EpCAM autoantibodies in a previous study [25]. All patients were in stage Duke's C or D. Seventy percent of the patients had metastatic disease with a high tumor burden, most patients received standard regimens of chemotherapy (5-FU with folinic acid, Oxaliplatin and/or Irinotecan) and/or radiotherapy. Unfortunately, it was only possible to collect a small number of samples from untreated patients to include in the analysis. In all cases samples were collected at a minimal interval of three weeks after the last chemotherapy application.

2.2. Monoclonal and polyclonal antibodies

Female BALB/c mice were immunized several times with purified recombinant EpCAM protein obtained as described [23]. Immunized animals serum contained >1 mg/ml polyclonal anti-EpCAM-immunoglobulin. Three days after the last boost spleen cells were fused with the myeloma cell line X63-Ag8.653 using polyethylene glycol 4000. Supernatants of growing hybridomas were screened by ELISA for reactivity with the EpCAM protein. Two hybridomas, EpCAM25.1 and EpCAM33.2 (both of IgG1 kappa isotype), were identified and cloned by limiting dilution. Mass production was performed in a MiniPERM modular bioreactor (Greiner, Frickenhausen, Germany) and immunoglobulin was purified by affinity chromatography over protein A Sepharose CL-4B (GE Healthcare, Freiburg, Germany). The specificity of purified mAbs was verified by Western blotting of lysate derived from Colo205 colon carcinoma cells. The mAbs M2-5 and M4-10 have been purchased from Biovendor (Heidelberg, Germany) and were generated in BALB/c mice with recombinant EpCAM protein produced from recombinant virus-infected silkworm larvae. Both have been described to bind to recombinant EpCAM as analyzed by BIACORE, in ELISA and Western blotting [17].

2.3. ELISA to determine soluble EGFII-EpCAM concentrations

EpCAM33.2 mAb (2.5 μg/ml) was coated in 20 mM carbonate-buffer (all reagents are from Sigma, Munich, Germany unless otherwise stated), pH 9.0 at 100 μl per well to Black Fluoro-Nunc MaxiSorb Plates (Nunc, Rochester, USA) at 4 °C overnight. After one wash with TPBS, 400 μl blocking buffer (2% high purity human serum albumin (Calbiochem, Darmstadt, Germany; highest purity), 0.1% Tween-20, 0.1% sodium azide, 1 mM EDTA in PBS, pH 7.4) was added. After an overnight incubation, the plates were washed four times with TPBS, and human sera (1:5 dilution in TPBS) were tested in duplicate or triplicate. Ten micrograms recombinant EpCAM-protein in TPBS were added per well for the standard curve and a geometric dilution was done to at least 1 ng. Because of the relatively high concentration of soluble EpCAM, we did not use any human serum for diluting the standard curve to prevent Ab interaction. The plates were incubated and washed as above and 100 μl TPBS with 250 ng biotinylated mAb EpCAM 25.1 was added to each well. The incubation was done for 2 h. After washing 100 μl 0.5 μg/ml Streptavidin-AKP (alkaline phosphatase. BD Pharmingen, Heidelberg, Germany) was added to each well and after a 2 h incubation at 4 °C and five wash steps, 200 μl substrate-solution (0.2 mM 4-methylumbelliferyl phosphate, 0.05 M $NaCO_3$, 2 mM $MgCl_2$) was added to each well and the plates were incubated for 3 h at room-temperature. The plates were measured in a Victor II Spectrophotometer (Wallac/PerkinElmer, Rodgau-Jügesheim, Germany) with an excitation filter at 365 nm and an emission filter at 450 nm. Every plate contained control wells to exclude variations between different plates.

2.4. ELISA to determine anti-EpCAM-immunoglobulins

The method has been described before [25]. Sera were added in a geometric dilution in TPBS (1:10 to 1:80).

2.5. Direct comparison of anti-MK-1 ELISA (untrimmed-EpCAM specific) with the mAbs in this study

60 randomly selected patient sera have been tested with the commercial untrimmed-EpCAM specific (Anti-MK-1) ELISA from Biovendor (according to Abe et al.) and in parallel with coated plates with Anti-EGFII-EpCAM clones 33.2 and 25.1 with a photometric detection reaction as provided by Biovendor. The protocol from Abe et al. was used also with the new mAbs described in this study. For the standard curve highly purified native folded EpCAM from S2-cells has been used.

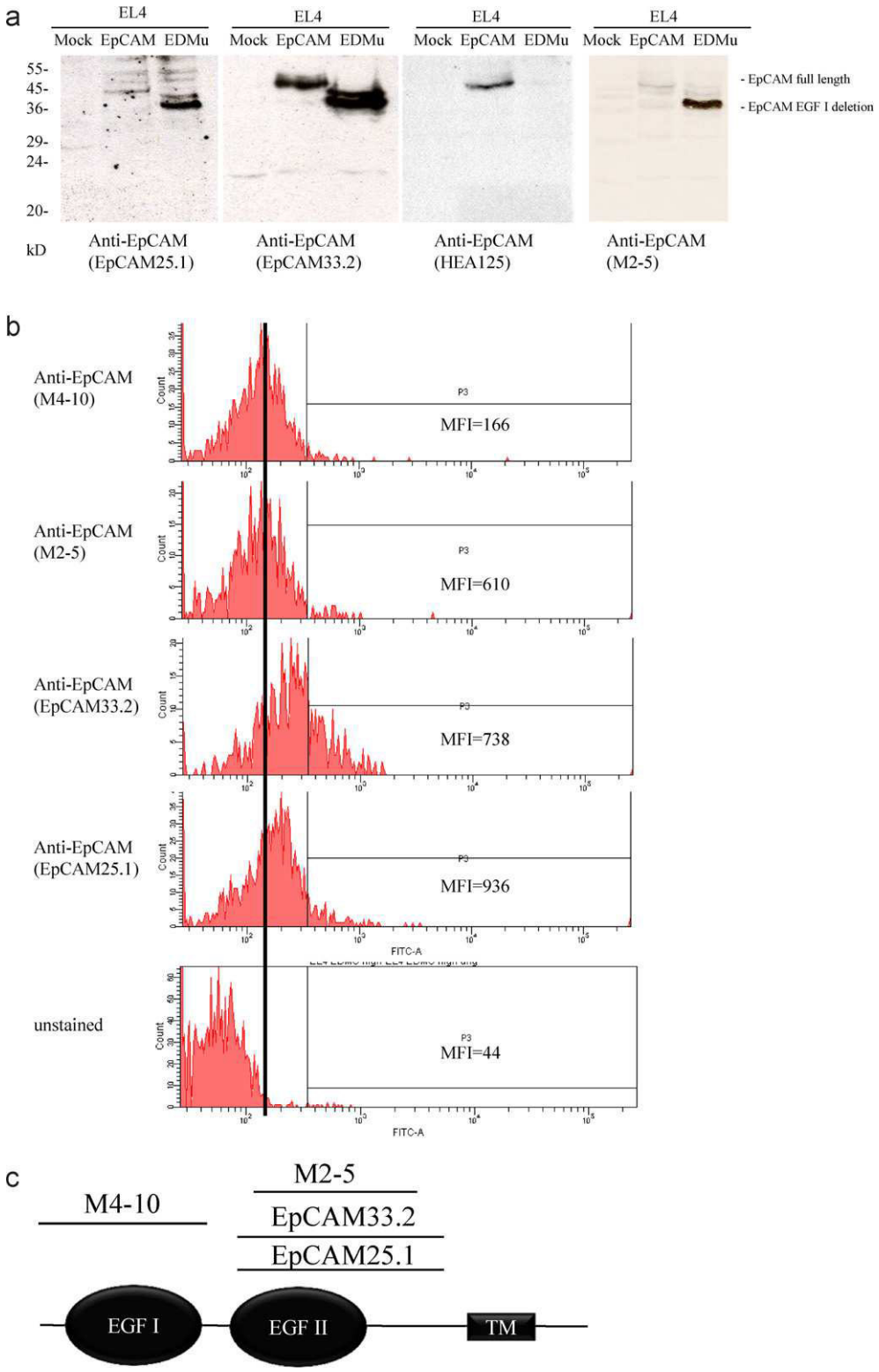

Fig. 1. (a) Binding of mAbs to full-length EpCAM or to an EGF-like domain I deletion mutant (EDMu) as analyzed by Western blotting. Hundred micrograms lysate from stably transfected cells has been used per lane. (b) Indirect staining of mAb binding to EL4 cells transfected with an EGF-like domain I deletion mutant (EDMu). The cells have been preincubated with the indicated mAbs and stained thereafter with directly labeled detector mAb. The mean-fluorescence index (MFI) is shown in the histograms. (c) Domain-specific binding of the mAbs in this study to the two major extracellular domains of untrimmed EpCAM is shown. This model is congruent with all epitope studies performed.

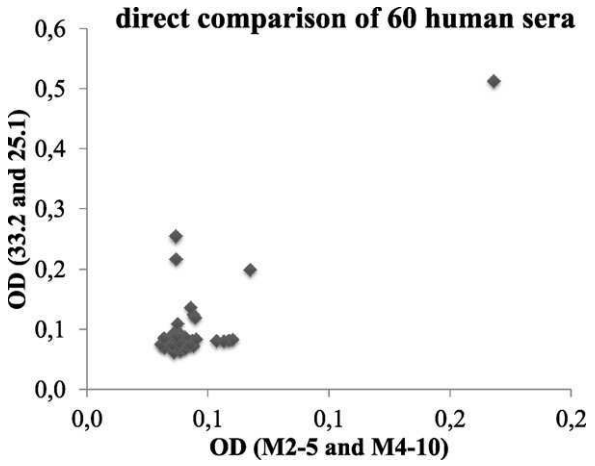

Fig. 2. 60 randomly selected patient sera have been tested with the commercial Anti-MK-1 ELISA from Biovendor (manufactured according to Abe et al.) and in parallel with coated plates with Anti-EpCAM 33.2 and 25.1 with a photometric detection reaction. The protocol from Abe et al. was used with the new mAbs described in this study.

3. Results

3.1. Development of the anti-EGFII-EpCAM-mAbs

Repeated immunization of mice with recombinant EpCAM-protein led to a vigorous humoral immune response. Two novel hybridomas named EpCAM25.1 and EpCAM33.2, both secreting an IgG1 kappa anti-EpCAM Ab were established. To determine the epitope specificity the antibodies were tested by Western blotting for binding to the full size EpCAM protein and to an EGF-like domain I deletion mutant (Fig. 1a). Both new mAbs reacted with full-length EpCAM and the deletion variant indicating that they recognize the second EGF-like repeat (Fig. 1).

3.2. Direct comparison of M2-5 and M4-10 with anti-EpCAM 33.2 and 25.1 Ab pairs in a photometric ELISA: untrimmed-EpCAM versus EGFII-EpCAM

To visualize differences in the EpCAM binding properties of both ELISA mAb pairs, we analyzed 60 patients' sera and a titration curve of recombinant standard protein (Fig. 2 and Supplementary Fig. 3). Some samples gave similar results with both assays; however more sera tend to result in a lower signal in the M2-5/M4-10 (untrimmed EpCAM) ELISA. Remarkably two sera were only positive in the EGFII-EpCAM-ELISA with Anti-EpCAM 33.2 and 25.1, and four sera were only positive untrimmed-EpCAM specific ELISA with M2-5 and M4-10. The titration curve showed a lower signal at higher EpCAM protein concentrations in the untrimmed-EpCAM specific ELISA without any signal increase in this range.

3.3. Development of the anti-EpCAM-EGFII-ELISA

We tested the various EpCAM mAbs and the polyclonal rabbit antiserum in all possible combinations in an ELISA. The best signal-to-noise ratios were obtained employing the new mAb EpCAM33.2 either in combination with EpCAM25.1 or with HEA125 as detection antibodies (Fig. 3a). The combination with HEA125 demonstrated a highly significant lower sensitivity at low EpCAM-concentrations (around 50 ng per well, Fig. 3b). Further optimizations (magnesium-concentrations, incubation time, substrate, plate type) allowed us to set an assay able to detect EpCAM serum concentrations in the range of 0.005–75 μg/ml (Fig. 4a and b, Supplementary Figs. 4 and 5). The signal-to-noise ratio was best at 180 min and stable over 1 h (data not shown). The detection limit of the test was below 100 pg (signal to noise-ratio of 37.2). We used a potential equation to calculate the EpCAM-concentration from the fluorescence signal, which fitted best to the plotted data points at physiological EpCAM concentrations (Fig. 4b).

3.4. Soluble EGFII-EpCAM concentrations in human sera

With this assay, we analyzed 471 sera obtained from 59 healthy donors and 412 patients with adenocarcinoma (Fig. 5a). The measurement was done in duplicates or triplicates and the mean standard deviation (SD) was 5.5% (85% of the samples had a SD of under 0.1). All tested sera showed detectable soluble EGFII-EpCAM (at least a tenfold higher signal than the detection limit).

The mean EpCAM concentration in all samples was 1915 ng/ml (Fig. 5a). In the sera from healthy controls ($N = 59$) we found an EpCAM-concentration ranging from 232 to 8893 ng/ml with a mean of 1525 ng/ml. In patients with adenocarcinoma we measured a concentration from 176 to 36,259 ng/ml with a mean of 1971 ng/ml. The mean values for the major disease groups were the following: colon cancer (mean 1313 ng/ml, $N = 60$), rectal cancer (mean 1923 ng/ml, $N = 86$), stomach cancer (mean 1392 ng/ml, $N = 34$), breast cancer (mean 2359 ng/ml, $N = 191$) and prostate cancer (mean 2013 ng/ml, $N = 32$).

The EpCAM concentrations did not correlate with the age of the donors (Fig. 5b). In addition, sera from 37 tumor-free patients with cardiovascular diseases did not show higher EpCAM-concentrations as compared to the healthy controls (data not shown).

Some rare patients (13 in total, which makes up 3.16% of the total number of patients in this study) had extremely high concentrations of soluble EGFII-EpCAM. Clinically there were no particular findings in these 13 patients concerning tumor stage and therapy.

3.5. Correlation of soluble EpCAM concentration with anti-EpCAM-immunoglobulin

In a previous study we had analyzed 560 serum samples for the presence of EpCAM-reactive Abs [25]. Four hundred and one of the stored, identical samples were analyzed again in this work and EpCAM protein concentration was determined as described above, while 181 new serum samples were simultaneously analyzed for both EGF-II-EpCAM protein concentration and EpCAM-reactive antibodies. IgG2, IgG3, IgG4 and partially IgA autoantibodies against EpCAM showed a negative correlation with the amount of circulating protein (Fig. 6): individuals with elevated levels of soluble EpCAM had much lower amounts of IgG2, IgG3 and IgG4 EpCAM autoantibodies and vice versa. A similar correlation was found for IgA, however more intermediate levels of antibodies and EpCAM were found. Anti-EpCAM IgM concentration did not show a positive or negative correlation with the soluble protein.

Only very low concentrations of EpCAM-reactive IgG1 were present (Fig. 6a). Only one of the 582 samples contained intermediate amounts of EpCAM protein, IgG2, IgG3 and IgG4 antibodies, most likely representing an outlier.

3.6. Reproducibility

To analyze intra-sample reproducibility, we tested 47 sera in two completely independent experiments (Supplementary Fig. 7a). In both tests similar values were obtained demonstrating a high reproducibility with a correlation factor R^2 of 0.96 after a linear equation was fitted to the data points. To determine inter-sample variability a second blood sample was obtained (minimal two weeks after the first sample) from 19 patients and tested

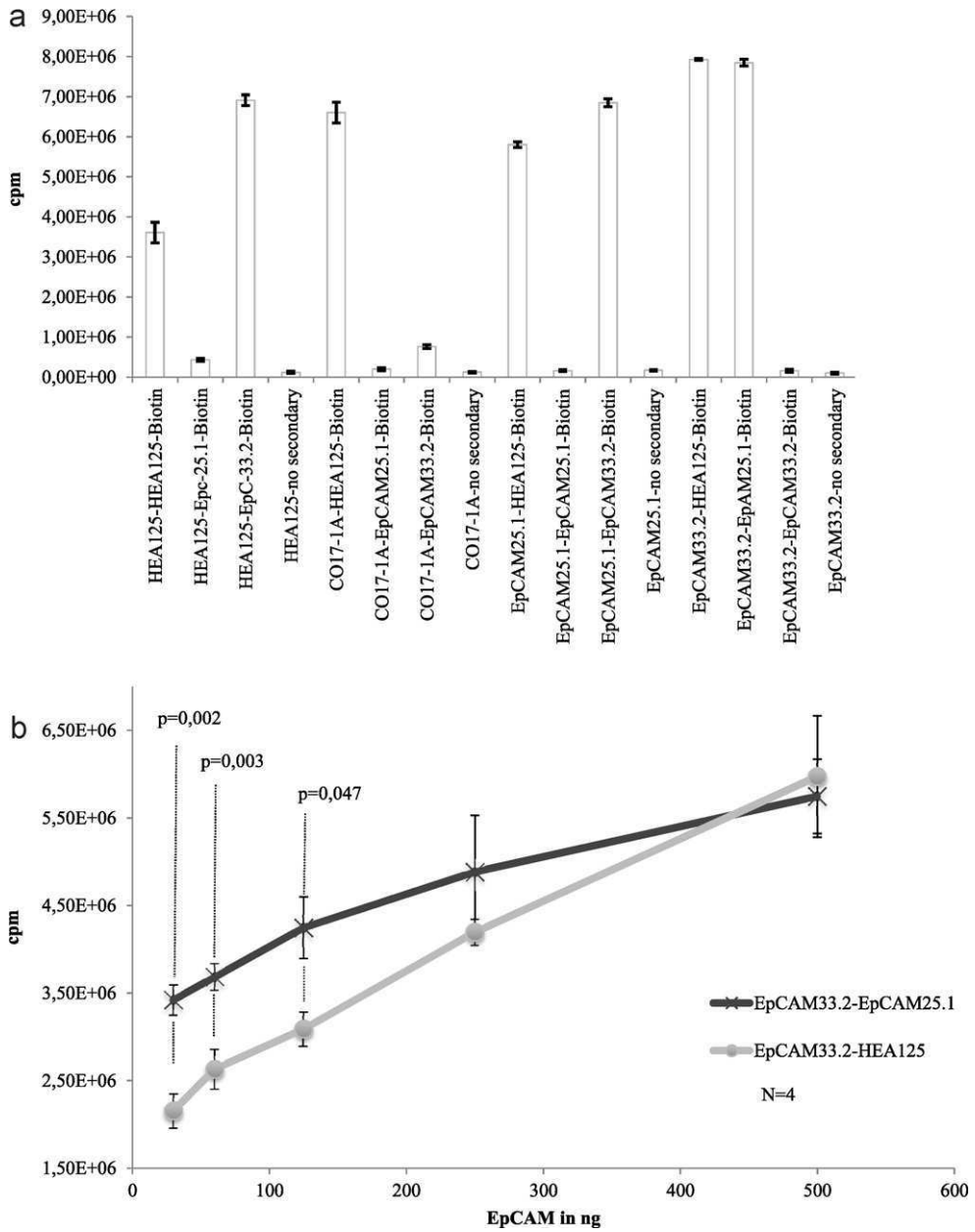

Fig. 3. (a) Optimization of ELISA-Ab pairs. The ELISA has been performed with 500 ng EpCAM-protein as sample and 180 min substrate incubation. Mean counts per minute (cpm) from six replicates and standard deviations are shown. (b) Direct comparison of the EpCAM33.2-EpCAM25.1 and EpCAM33.2-HEA125 ELISA Ab pairs. The EpCAM-protein has been titrated in a geometric dilution starting with 500 ng. Mean counts per minute (cpm) and standard deviations after 180 min substrate incubation from four replicates are shown. *p*-Values have been calculated with a paired Student's *T*-test.

(Supplementary Fig. 7b). The mean standard deviation of both analyses was 7.6%.

3.7. EpCAM immunocomplexes and low cytometric shift assays

To exclude that immunocomplexed EpCAM might prevent detection by the ELISA antibodies, we immunoprecipitated Ab from sera containing high amounts of anti-EpCAM IgG or high amounts of soluble EpCAM. Sera with at least 10 ng/ml anti-EpCAM-antibodies were tested. We could not identify any soluble, Ab-bound EpCAM in 42 sera with the highest ELISA-titers by western blotting after immunoprecipitation (Supplementary Fig. 8). The detection limit was below 25 ng. Even in sera with high Anti-EpCAM-IgG as determined by ELISA, no binding of Ab to Colo205 could be detected by flow cytometry. However a discrete inhibition of HEA125-FITC binding to Colo205 could be noted (Supplementary Figs. 9 and 10).

4. Discussion

In this study we have found high levels of circulating EpCAM in healthy controls and patients. Considering the large amount of literature on EpCAM, with more than hundred clinical studies published including several immunotherapeutic trials it is rather surprising that so far only three reports have dealt with the issue of circulating EpCAM protein in the serum and none analyzed the different protein isoforms [16–18].

Abe and Petsch have reported amounts of circulating EpCAM two to three log scales lower than those found in the present study. According to Abe et al., both healthy controls and 90% of the patients

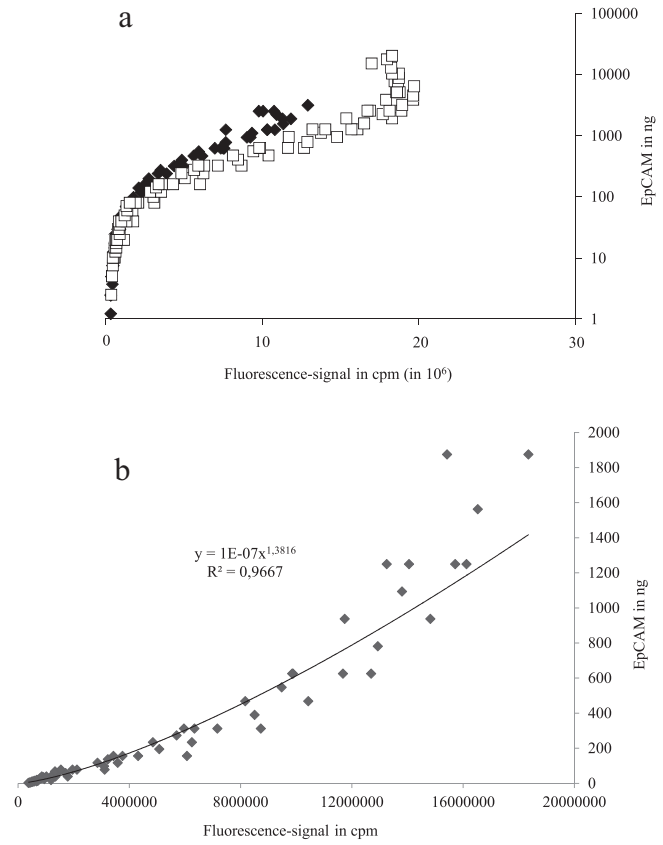

Fig. 4. (a) Data points from EpCAM standard curves. Data points from two geometric dilutions, ranging from 1 ng to 50 μg recombinant EpCAM (log-scale) are shown. The dilutions and measurements were done in two completely different experiments demonstrating the reproducibility and the large measurement range of the fluorometric anti-EpCAM-ELISA. (b) Data points from EpCAM standard curves with low EpCAM concentrations. Data points from geometric dilutions of EpCAM-antigen with a potential fit are shown. The coefficient of determination (R^2) demonstrates a very good correlation of the equation used to calculate EpCAM-amounts in human sera.

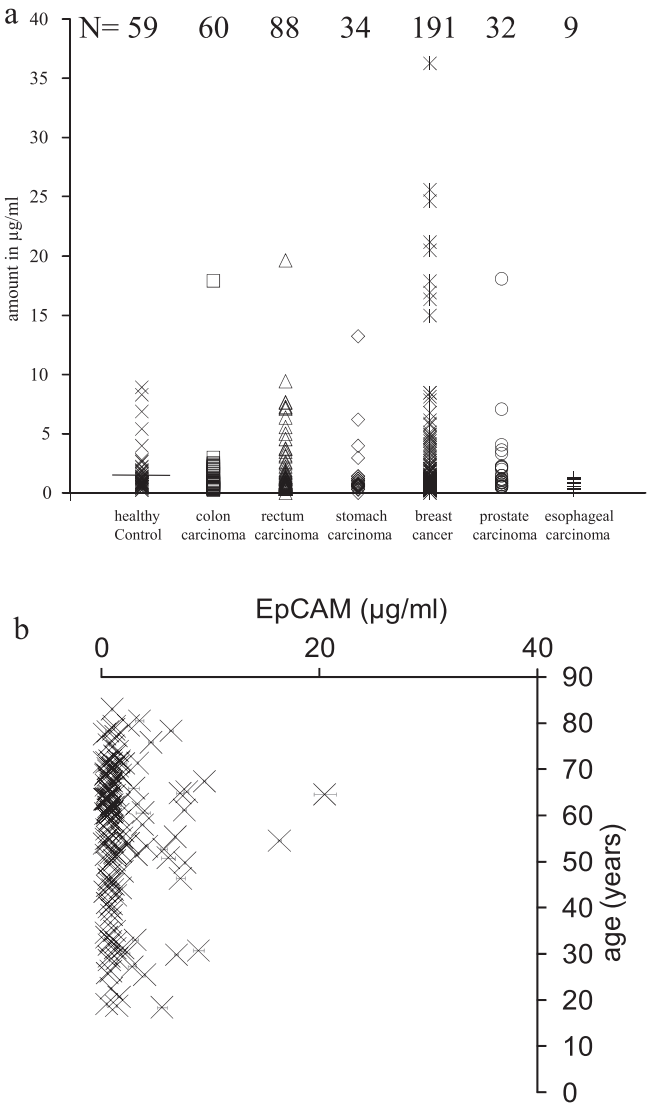

Fig. 5. (a) Soluble EpCAM protein in patients and healthy controls as measured by ELISA. Frozen sera have been diluted 1:5 with TPBS and were directly applied on the ELISA wells. The concentrations shown have been back calculated from the dilution. Shown are 59 sera from healthy controls (×), 104 sera from patients with colon carcinoma (□), 87 sera from patients with rectum carcinoma (△), 39 sera from patients with stomach carcinoma (◊), 281 sera from patients with breast cancer (*), 32 sera from patients with prostate carcinoma (○) and 5 sera from patients with esophageal cancer (+). Mean values from double-measurements are shown. (b) EpCAM-protein concentrations are blotted against age. Mean values and standard deviations from double-measurement of 153 randomly selected sera are shown.

had soluble EpCAM, in a concentration lower than 2 ng/ml. Ten percent of the patients showed increased amounts of EpCAM protein (in the range of 2–78 ng/ml) in the blood. In contrast Petsch et al. found detectable EpCAM in 17% of cancer patients, with concentrations one log scale lower than Abe [16].

Differences specificities of the mAbs used can explain the apparently much higher sensitivity of our assay and the differences in the direct comparison of the ELISA. We used mAbs which recognize two different epitopes of the second EGF-like repeat in the extracellular portion of EpCAM in order to minimize competition between the antibodies and to reduce dimerization artifacts. In contrast one mAb used by the other groups recognized the first EGF-like domain while the second recognized the second one. Therefore only full-length shed EpCAM can be detected in this ELISA.

According to the instruction manual of the ELISA from Abe et al., a 1:1 serum dilution was used which might lead to serum inhibition of the ELISA. Moreover, bovine milk protein (Block-Ace) had been used for blocking of the ELISA plates, the standard protein was denaturated and its glycosylation pattern was insect-cell congruent. We have reported earlier that this might disturb Ab binding to EpCAM and hence might lead to lower detection of EpCAM-levels [25]. The commercially available ELISA kit based on the method from Abe et al. has been discontinued by the company Biovendor. An ELISA plate was still available to enable us a direct comparison. We found similar results in some of the sera, yet 10% of the samples were single-positive only in one of the ELISAs. Because different soluble forms of the EpCAM protein have been described, the most likely explanation is that some forms are only recognized by some antibodies depending on the epitope. Because the antibodies described in the present study were generated against native folded protein, while the M2-5 and M4-10 were generated with partially denaturated protein, it might also be possible that M2-5 and M4-10 recognize missfolded parts of the protein released e.g. after tumor necrosis. Remarkably the M2-5/M4-10 based ELISA was not able to bind higher amounts of EpCAM and might therefore miss EpCAM in the physiological range without high dilution. In addition to the underestimation due to different standard protein, this might be another explanation why lower soluble EpCAM levels measured with this ELISA have been described.

Petch et al. used a polyclonal, commercial goat anti-EpCAM-serum to coat ELISA plates and detected reactivity by adding a mAb.

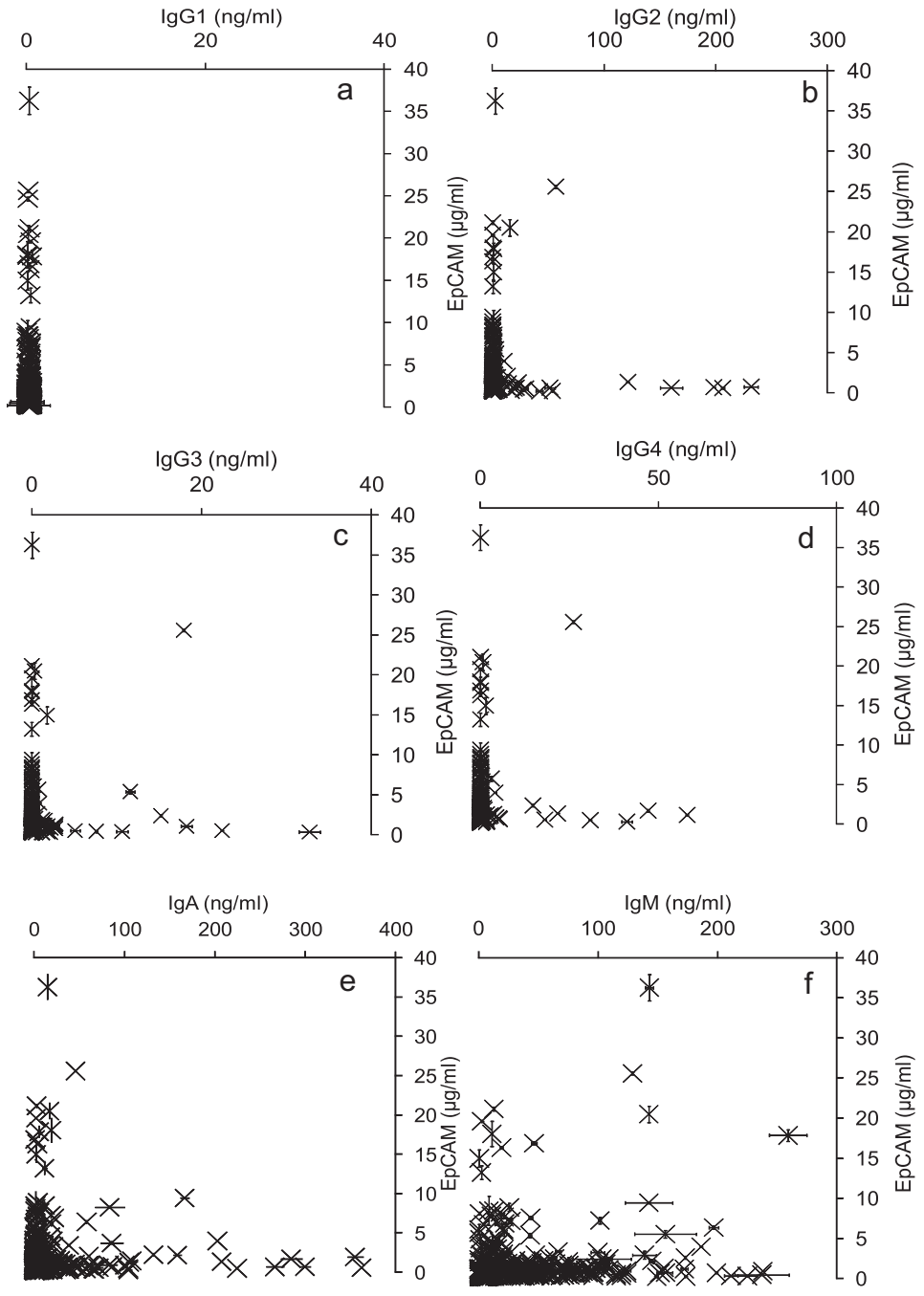

Fig. 6. (a–f) Correlation of EpCAM-reactive Ab levels and soluble EpCAM protein levels. Shown are 582 sera, subtype-specifically analyzed for EpCAM-reactive IgG1-, IgG2-, IgG3-, IgG4-, IgM- and IgA-autoantibodies (data from our last study with additional samples measured in this study), and for soluble EpCAM (done for this study). Mean values and standard deviations from double-measurements are shown.

In our opinion, sensitivity and specificity might be lower when using a polyclonal serum for coating as compared to the mAb we used. They identified shed EpCAM in the supernatant of all tested EpCAM-positive cancer cell lines but only low concentrations in the sera.

Kimura et al. used the commercial available ELISA kit from Biovendor to determine EpCAM-concentration pre-operative in sera from patients with esophageal cancer. The authors did not use the standard protein in the kit and therefore only measured the EpCAM concentration in U/ml. They showed that the EpCAM-level in 60 sera from patients with esophageal cancer is significantly increased as compared to 20 sera from healthy controls ($p = 0.02$). They found no significant correlation between the clinicopathologic characteristics of esophageal cancer patients and serum EpCAM levels in the serum. However the survival rates of patients with high EpCAM levels was significantly better as compared to those with low EpCAM-amounts (5-year survival rate 75.8% vs. 44.5%.).

We identified an increased amount of EGFII-EpCAM protein in only 2% of the patients with adenocarcinoma. This is not much different from the percentage of patients that have circulating autoantibodies to native EpCAM according to our previous findings (around 4.5%) [25], which is also lower than what had been reported in other publications.

High amounts of EGFII-EpCAM protein in the serum were consistently associated with low amounts of EpCAM-reactive IgG2, IgG3, IgG4 and IgA antibodies. Similar correlations have been described for other proteins and the respective autoantibodies [26–28].

The concentrations of the EGFII-EpCAM protein found in our study are remarkably high (in a microgram-per-milliliter range). This is not surprising, as EpCAM is expressed in almost all complex epithelial tissues. Cleavage of EpCAM is sequentially catalyzed by TACE followed by releasing its ectodomain EpEX [15]. Interestingly, cleavage of the extracellular domain EpEX seems to be a prerequisite for subsequent intramembrane cleavage by presenilin and release of EpICD, which together with FHL2 creates a "EpCAM signalome" which rapidly upregulates expression of c-myc, Cyclin E and e-fabp, and induces cell proliferation [29]. Therefore release of EpEx might me an important step in maintaining normal epithelial cell proliferation and therefore regeneration of the GI mucosa and the skin. The concentrations (μg/ml) and cut-off of our EGFII-EpCAM-ELISA are also comparable to other ELISAs measuring colon-cancer antigens [30].

EpCAM has also been shown to be present in tumor cell-derived microvesicles (TMV) which are shed from colon cancer cells and taken up by monocytes [31]. Potentially EpCAM could be released from these TMV or from the endosomes of monocytes in soluble form.

An exact determination of circulating EpCAM isoforms and of EpCAM autoantibodies in sera might be severely affected by the presence of immune complexes. Missed detection of antigen bound in immune complexes is a possible drawback of ELISA methods. A highly sensitive Western blot was therefore used to search for IgG-bound EpCAM. The detection limit of our Western blot was below 25 ng EpCAM, which corresponds to about 0.5 ng/μl serum. With this detection limit we should be able to detect EpCAM–immune complexes in the sera, if only about 2% of the soluble EpCAM would be involved in immune complex formation. This makes it extremely unlikely, that our ELISA is not detecting circulating EpCAM or EpCAM autoantibodies because of immune complex formation interference.

Soluble, glycosylated EpCAM seems to be extremely resistant to degradation and proteolytic cleavage: it remains native at pH 7.0 up to 60 °C (Supplementary Figs. 1 and 2). Accordingly, a long plasma-half-life of EGFII-EpCAM could explain the high concentrations we found as the second EGF-like domain is responsible for homodimerization.

Novel promising bi- and trifunctional antibodies targeting EpCAM have been developed [32,33]. A recent study showed that higher amounts of soluble EpCAM can block the function of the bifunctional antibodies. Cytotoxicity of the therapeutic Ab was abolished at EpEX concentrations of 10 μg/ml [16]. According to our measurements, tumor-reactive cytotoxicity would be reduced to 50% in the majority of patients. Due to the oligomerization of soluble EpCAM, EpCAM/CD3-bispecific antibodies could be "cross-linked" in human plasma, inducing CD3-clustering on the T-cell-surface with following T cell activation which might explain the potentially dangerous clinical side effects which have been reported. Indeed, CD69-upregulation was induced with 30 ng/ml EpEX or more [16].

In this study we prove the presence of soluble EpCAM and detected its EGF-II-repeat containing isoform in a higher concentration as previously published for untrimmed EpCAM. In light with recent publications, this high EGFII-EpCAM concentration will reduce the clinical activity of novel EpCAM-targeting bispecific antibodies and increase side effects. With better characterization of the soluble EpCAM present in human plasma, development of bispecific antibodies which target only cell-bound forms will greatly increase clinical response and reduce side effects.

Acknowledgments

The authors wish to thank Thomas Blankenstein, Wolfgang Uckert (both MDC, Berlin) and Wolfgang Zimmermann (Univ. of Munich) for helpful discussion. This work was supported by the Max-Delbrück-Centrum for Molecular Medicine (part of the Helmholtz association) and the Charité University Medicine Berlin.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.imlet.2012.02.004.

References

[1] Herlyn M, Steplewski Z, Herlyn D, Koprowski H. Colorectal carcinoma-specific antigen: detection by means of monoclonal antibodies. Proc Natl Acad Sci USA 1979;76:1438.
[2] Linnenbach AJ, Wojcierowski J, Wu SA, Pyrc JJ, Ross AH, Dietzschold B, et al. Sequence investigation of the major gastrointestinal tumor-associated antigen gene family, GA733. Proc Natl Acad Sci USA 1989;86(1):27.
[3] Szala S, Froehlich M, Scollon M, Kasai Y, Steplewski Z, Koprowski H, et al. Molecular cloning of cDNA for the carcinoma-associated antigen GA733-2. Proc Natl Acad Sci USA 1990;87(9):3542.
[4] Riethmuller G, Holz E, Schlimok G, Schmiegel W, Raab R, Hoffken K, et al. Monoclonal antibody therapy for resected Dukes' C colorectal cancer: seven-year outcome of a multicenter randomized trial. J Clin Oncol 1998;16(5):1788.
[5] Moldenhauer G, Momburg F, Möller P, Schwartz R, Hämmerling GJ. Epithelium specific surface glycoprotein of Mr 34,000 is a widely distributed human carcinoma marker. Br J Cancer 1987;56:712.
[6] Momburg F, Moldenhauer G, Hammerling GJ, Moller P. Immunohistochemical study of the expression of a Mr 34000 human epithelium-specific surface glycoprotein in normal and malignant tissue. Cancer Res 1987;47:2883.
[7] Baeuerle PA, Gires O. EpCAM (CD326) finding its role in cancer. Br J Cancer 2007;86:417.
[8] Kuhn SKM, Nübel T, Ladwein M, Antolovic D, Klingbeil P, Hildebrand D, et al. A complex of EpCAM, claudin-7, CD44 variant isoforms, and tetraspanins promotes colorectal cancer progression. Mol Cancer Res 2007;5(6):553.
[9] Yamashita T, Forgues M, Wang W, Kim JW, Ye Q, Jia H, et al. EpCAM and alpha-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma. Cancer Res 2008;68(5):1451.
[10] Balzar M, Winter MJ, de Boer CJ, Litvinov SV. The biology of the 17-1A antigen (Ep-CAM). J Mol Med 1999;77(10):699.
[11] Balzar M, Briaire-de Bruijn IH, Rees-Bakker HA, Prins FA, Helfrich W, de Leij L, et al. Epidermal growth factor-like repeats mediate lateral and reciprocal interactions of Ep-CAM molecules in homophilic adhesions. Mol Cell Biol 2001;21(7):2570.
[12] Thampoe IJ. NJSaLKO: biochemical analysis of a human epithelial surface antigen: differential cell expression and processing. Arch Biochem Biophys 1988;15(267):342.
[13] Gires O. EpCAM a proteolytically cleaved oncogene and an excellent therapeutic target in cancer. Med Sci (Paris) 2009;25(5):449.
[14] Denzel S, Maetzel D, Mack B, Eggert C, Barr G, Gires O. Initial activation of EpCAM cleavage via cell-to-cell contact. BMC Cancer 2009;9:402.
[15] Maetzel D, Denzel S, Mack B, Canis M, Went P, Benk M, et al. Nuclear signalling by tumour-associated antigen EpCAM. Nat Cell Biol 2009;11(2):162.
[16] Petsch S, Gires O, Ruttinger D, Denzel S, Lippold S, Baeuerle PA, et al. Concentrations of EpCAM ectodomain as found in sera of cancer patients do not significantly impact redirected lysis and T cell activation by EpCAM/CD3-bispecific BiTE antibody MT110. MAbs 2011;3(1).
[17] Abe H, Kuroki M, Imakiire T, Yamauchi Y, Yamada H, Arakawa F, et al. Preparation of recombinant MK-1/Ep-CAM and establishment of an ELISA system for determining soluble MK-1/Ep-CAM levels in sera of cancer patients. J Immunol Methods 2002;270(2):227.
[18] Kimura H, Kato H, Faried A, Sohda M, Nakajima M, Fukai Y, et al. Prognostic significance of EpCAM expression in human esophageal cancer. Int J Oncol 2007;30(1):171.
[19] Rodriguez CR, Fei DT, Keyt B, Baly DL. A sensitive fluorometric enzyme-linked immunosorbent assay that measures vascular endothelial growth factor165 in human plasma. J Immunol Methods 1998;219(1–2):45.
[20] Gutierrez J, Konecny GE, Hong K, Burges A, Henry TD, Lambiase PD, et al. A new ELISA for use in a 3-ELISA system to assess concentrations of VEGF splice variants and VEGF(110) in ovarian cancer tumors. Clin Chem 2008;54(3):597.
[21] Konecny GE, Meng YG, Untch M, Wang HJ, Bauerfeind I, Epstein M, et al. Association between HER-2/neu and vascular endothelial growth factor expression predicts clinical outcome in primary breast cancer patients. Clin Cancer Res 2004;10(5):1706.
[22] Plumer A, Duan H, Subramaniam S, Lucas FL, Miesfeldt S, Ng AK, et al. Development of fragment-specific osteopontin antibodies and ELISA for quantification in human metastatic breast cancer. BMC Cancer 2008;8:38.

[23] Anborgh PH, Wilson SM, Tuck AB, Winquist E, Schmidt N, Hart R, et al. New dual monoclonal ELISA for measuring plasma osteopontin as a biomarker associated with survival in prostate cancer: clinical validation and comparison of multiple ELISAs. Clin Chem 2009;55(5): 895.
[24] Gonzalez A, Alegre E, Arroyo A, LeMaoult J, Echeveste JI. Identification of circulating nonclassic human leukocyte antigen G (HLA-G)-like molecules in exudates. Clin Chem 2011;57(7):1013.
[25] Schmetzer O, Moldenhauer G, Riesenberg R, Pires JR, Schlag P, Pezzutto A. Quality of recombinant protein determines the amount of autoreactivity detected against the tumor-associated epithelial cell adhesion molecule antigen: low frequency of antibodies against the natural protein. J Immunol 2005;174:942.
[26] Nossal GJ, Pike BL. Evidence for the clonal abortion theory of B-lymphocyte tolerance. J Exp Med 1975;141:904.
[27] Karvelas M, Nossal GJ. Memory cell generation ablated by soluble protein antigen by means of effects on T- and B-lymphocyte compartments. Proc Natl Acad Sci USA 1992;89:3150.
[28] Nossal GJ, Karvelas M. Soluble antigen abrogates the appearance of anti-protein IgG1-forming cell precursors during primary immunization. Proc Natl Acad Sci USA 1990;87:1615.
[29] Münz MKC, Mack B, Schmitt B, Zeidler R, Gires O. The carcinoma-associated antigen EpCAM upregulates c-myc and induces cell proliferation. Oncogene 2004;23(34):5748.
[30] Leman ES, Schoen RE, Weissfeld JL, Cannon GW, Sokoll LJ, Chan DW, et al. Initial analyses of colon cancer-specific antigen (CCSA)-3 and CCSA-4 as colorectal cancer-associated serum markers. Cancer Res 2007;67(12):5600.
[31] Baj-Krzyworzeka M, Szatanek R, Weglarczyk K, Baran J, Urbanowicz B, Branski P, et al. Tumour-derived microvesicles carry several surface determinants and mRNA of tumour cells and transfer some of these determinants to monocytes. Cancer Immunol Immunother 2006;55(7):808.
[32] Salnikov AV, Groth A, Apel A, Kallifatidis G, Beckermann BM, Khamidjanov A, et al. Targeting of cancer stem cell marker EpCAM by bispecific antibody EpCAMxCD3 inhibits pancreatic carcinoma. J Cell Mol Med 2009;13(9B):4023.
[33] Sebastian M, Kuemmel A, Schmidt M, Schmittel A. Catumaxomab: a bispecific trifunctional antibody. Drugs Today (Barc) 2009;45(8):589.

Peptide Binding at Class I Major Histocompatibility Complex Scored with Linear Functions and Support Vector Machines

Henning Riedesel, Björn Kolbeck, Oliver Schmetzer, Ernst-Walter Knapp
riedesel@chemie.fu-berlin.de, bjko@chemie.fu-berlin.de, o.schmetzer@mdc-berlin.de, knapp@chemie.fu-berlin.de

Institute of Chemistry, Free University of Berlin, Takustrasse 6, 14195 Berlin, Germany

Abstract

We explore two different methods to predict the binding ability of nonapeptides at the class I major histocompatibility complex using a general linear scoring function that defines a separating hyperplane in the feature space of sequences. In absence of suitable data on non-binding nonapeptides we generated sequences randomly from a selected set of proteins from the protein data bank. The parameters of the scoring function were determined by a generalized least square optimization (LSM) and alternatively by the support vector machine (SVM). With the generalized LSM impaired data for learning with a small set of binding peptides and a large set of non-binding peptides can be treated in a balanced way rendering LSM more successful than SVM, while for symmetric data sets SVM has a slight advantage compared to LSM.

Keywords: major histocompatibility complex, peptide binding, separating hyperplane, support vector machine, learning and predicting, scoring function.

1 Introduction

Every adaptive immune reaction is based on the specific detection of foreign substances by lymphocytes. These lymphocytes than destroy infected cells and/or stimulate an antibody response, which generally leads to the complete removal of an invading microorganism from the body. Absolutely essential for such a successful immune response is the presentation of the foreign substances, which are in most cases peptides derived for instance from a replicating virus. These peptides are generated from the proteasom and transported to the endoplasmatic reticulum, where they are loaded in the major histocompatibility complex (MHC) (1-3). This complex together with the peptide is transferred to the cell surface and can be recognized by T-cells via the T-cell-receptor (TCR) (4, 5). Without presentation of peptides, no immune response against viruses can be initiated which leads to death of the organism and is the strategy of many pathogens (6). Not all peptides can be presented in the MHC. The binding depends on so-called anchor-amino-acids, which bind often with low specificity to the MHC, leaving the residual peptide exposed to the TCR (7, 8).

The development of vaccines, immunotherapies and the understanding of a pathogen crucially depend on the know-ledge of the immuno-dominat peptides from a target organism. Identification of these peptides can be done by binding assays in vitro after all possible peptides have been synthesized (9, 10). This is an extremely expensive approach, because even a very small virus encodes a considerable number of medium size proteins. For each of these proteins hundreds of peptides have to be synthesized and their ability to bind at the MHC must be probed in experiment. This often shows that only very few peptides can indeed bind to the MHC and that from thousands of screened peptides only one or two bind with high affinity, which is required for a

functional immune response.

To simplify the search for immuno-dominant peptides, several groups collected data of peptides that bind at MHC to generate a database, which can serve as starting point of computer-based methods to predict the ability of peptides to bind at MHC in silico. These approaches can help to reduce the number of peptides, which have to be tested in vitro. The most often used database of MHC binding peptides is the SYFPEITHI-database (SYF) (11). Another database for MHC binding peptides that offers however no prediction scheme is MHCPEP (PEP) (12). The SYF database refers to published data only. It contains about 3,500 MHC binding peptides, which are natural ligands to T-cell epitopes. The MHCPEP database is with about 13,000 MHC binding peptides considerably larger than SYF but may be less reliable, since it allows also for direct submission of data.

Generally, there are sequence based and structure based approaches to predict the ability of peptides to bind at the MHC. The latter uses X-ray structures of MHC or even better of the MHC-peptide complex as a starting point to model the binding geometry of different peptides (13). The structure based approach has the advantage to require only knowledge of one or at most a few crystal structures to study the peptide binding and provides a deeper understanding of the importance of specific interactions between peptides and the MHC. For peptides that bind to the MHC HLA-A*0201 it is evident from crystal structures that the binding peptides are often nonamers, which possess typically two hydrophobic anchor residues Lys at position 2 and Val at position 9 (14). This knowledge was a starting point to design empirical scoring functions that use also informations from sequence databases (11). However, a disadvantage of the structure based approach is the difficulty to estimate an error margin.

More recently a number of theoretical groups have employed bioinformatic methodology to predict the ability of peptides to bind at MHC based on sequence information. Among these methods are neural networks (15, 16) and methods based on scoring functions that are optimized by least square fitting (17) or by using the support vector machine (18, 19). In this study, we tried to explore a method to predict immuno-dominant epitopes by using a linear scoring function in sequence space.

2 Method

Peptide data bases. For the set of polypeptide sequences that bind at the MHC, we considered the SYF (11) and PEP (12) data bases in September 2003. There are 268 peptides in SYF that bind to the MHC HLA-A*0201. From these sequences 204 possess the canonical length of 9 residues. The remaining 64 peptides possess sequences longer than 9 residues. The peptides in SYF are sequence aligned i.e. equivalent sequence positions of different peptides were identified such that the corresponding peptide residues are supposed to interact with same residues of the MHC binding groove. We used this information to cut the length of the peptides longer than 9 residues to obtain nonapeptides. In the PEP database there are 506 peptides that bind to the MHC HLA-A*0201. These peptides were not aligned. Therefore, we considered only the nonapeptides. We merged these two sets of nonapeptides, which after removing the identical peptides yielded a database $\mathbb{S}^+$ of 538 nonapeptides binding at the MHC HLA-A*0201.

There are no explicit non-binding peptides available. We assumed that randomly chosen nonapeptides are unlikely to bind at the MHC. Hence, we generated up to 10,000 different nonapeptides $\mathbb{S}^-$ that were randomly taken from the concatenated sequences of 202 proteins selected from the protein database (20) (Table 1). Care was taken that the selected proteins do not contain nonapeptides that bind at MHC, although this can not be excluded.

Data representation. We assume that two sets of polypeptide sequences are available: one set of binding peptides $\mathbb{S}^+ = \{\vec{x}_n^+,\ n = 1, \ldots N^+\}$ and one set of non-binding peptides $\mathbb{S}^- = \{\vec{x}_n^-,\ n = 1, \ldots N^-\}$, which are obtained as explained above. For the present application all sequences considered are aligned and of equal length say M = 9. The polypeptide sequences $\vec{x}_n$ are represented by M subvectors

$$\vec{x}_n^t = (\vec{x}_{1,n}^t, \vec{x}_{2,n}^t, \ldots \vec{x}_{M,n}^t)\ , \tag{1}$$

where each subvector in (1) possesses 20 components

$$\vec{x}_{m,n}^t = (x_1^{(m,n)}, x_2^{(m,n)}, \ldots x_{20}^{(m,n)})\ , \tag{2}$$

which refer to the 20 native amino acids. Note that the superscript t refers to a row vector representation.

Individual components of a sequence vector $\vec{x}_n$ denoting the occurrence of amino acid type *j* at sequence position *m* will be addressed as $(\vec{x}_n)_{jm}$. The amino acid type at a particular sequence position is coded by setting the corresponding component of the subvector to unity, while all other 19 components of this subvector contain zero. Thus, from a more general view point the components of each subvector can also be interpreted as a probability distribution to find specific amino acid types at the corresponding sequence position. This interpretation becomes more meaningful, when averages $\langle\vec{x}\rangle$ of those sequence vectors are considered as is done below.

Table 1

PDB [a] codes of proteins whose concatenated sequences were used to generate the non-binding nonapeptides

1	**7ZNF**	**1AGQ**	**1BRX**	**1C51**	**1BCC**	**1EPW**	**1A75**	**1E4T**	**1AQU**	**1O23**	**1ISN**	**2IAD**
2	**6RLX**	**1AIR**	**8TIM**	**1BQP**	**1P3H**	**1E1H**	**1LF4**	**1E3E**	**1ALB**	**1NMM**	**1IQ1**	**2DLF**
3	**6Q21**	**1AFO**	**1A0R**	**1EIS**	**1UJL**	**1DXR**	**1E0C**	**1DX1**	**1AI1**	**1NCI**	**1IKN**	**1SUH**
4	**1HNE**	**3MRA**	**1A12**	**1C01**	**1GWY**	**1GWC**	**1HRK**	**1DSV**	**1AFV**	**1NAS**	**1IG3**	**1HA7**
5	**1EAD**	**1AUN**	**3ZNC**	**1C4R**	**1RK4**	**1G6R**	**1RIE**	**1DJ2**	**1A2Y**	**1N9P**	**1IFQ**	**1GK8**
6	**1VMO**	**1AUV**	**1A38**	**1GGX**	**1QKK**	**1FYT**	**1UOY**	**1DF3**	**1A1H**	**1MNU**	**1IFA**	**1BX7**
7	**821P**	**1A04**	**2SQC**	**1FHF**	**1B9C**	**1L0X**	**1H1V**	**1DD7**	**1914**	**1MBY**	**1IAL**	**1BWK**
8	**1BOM**	**1AF6**	**1BUG**	**1H4Y**	**1BFA**	**1ITZ**	**1GL5**	**1CQZ**	**1R2A**	**1MBE**	**1I7W**	**1BK6**
9	**1AHL**	**1A06**	**1BYO**	**1BPO**	**1BD2**	**1IR1**	**1GL2**	**1CL7**	**1QLX**	**1M4M**	**1I7E**	**1BJT**
10	**1SRA**	**1AXM**	**7PCK**	**1AB1**	**1AO7**	**1LFJ**	**1G74**	**1CE6**	**1PA2**	**1M3V**	**1I6Z**	**1ASZ**
11	**1DOX**	**1AZD**	**1BYY**	**1H8P**	**1A2X**	**1OM0**	**1FWU**	**1CDK**	**1P8J**	**1LB1**	**1I07**	**1AI9**
12	**1MSP**	**1AIW**	**2VSG**	**1GRW**	**1A2C**	**1OED**	**1FRB**	**1C2B**	**1ORS**	**1KCM**	**1HQV**	**1A6R**
13	**1FAT**	**1A0D**	**1B10**	**1JV1**	**1D9K**	**1TCR**	**1FKW**	**1BLN**	**1OMX**	**1KBQ**	**1HQ8**	**1A4H**
14	**1BGK**	**1BB9**	**1C3A**	**1O7N**	**1CNE**	**1QF3**	**1F93**	**1BKX**	**1OKQ**	**1K2F**	**1HN3**	**1A48**
15	**7UPJ**	**1A05**	**1EG5**	**1H0H**	**1CJK**	**1PJU**	**1F81**	**1BGX**	**1OGP**	**1JJO**	**1H96**	**1A2V**
16	**6UPJ**	**1BA1**	**1EHD**	**1B8M**	**1FV3**	**1WGT**	**1EDH**	**1BBS**	**1OCP**	**1JI9**	**2ZNC**	
17	**5UPJ**	**1BKD**	**1BQF**	**1GDJ**	**1EZF**	**1BR1**	**1E4W**	**1AX8**	**1OAA**	**1IWE**	**2MSS**	

[a] Ref. (20)

Scoring function. The decision that a sequence $\vec{x}$ is capable to bind or not, is performed with a scoring function, $f(\vec{x})$, which is linear in the sequences space $\mathbb{S}$ (feature space). The most general expression for the linear scoring function $f(\vec{x})$ is the linear form

$$f(\vec{x}) = \vec{w}^t \bullet \vec{x} + b \,, \tag{3}$$

where $\vec{x} \in \mathbb{S}$ is a 20*M component vector characterizing a particular sequence, $\vec{w}^t$ is a row vector of the same dimension as $\vec{x}$ and b is a scalar. The 20*M + 1 free parameters of the scoring function $\vec{w}^t$ and b are determined for a set of sequences, the so called learning set $\mathbb{S}_{learn}$ such that $f(\vec{x})$ adopts a value close to +1 for the binding sequences and close to –1 for the non-binding sequences. Hence, setting $f(\vec{x}) = 0$ defines a hyperplane in the 20*M dimensional sequence space $\mathbb{S}$ with plane normal vector $\vec{w}$. The hyperplane separates binding sequences $\vec{x}^+$ with $f(\vec{x}^+) > 0$ from non-binding sequences $\vec{x}^-$ with $f(\vec{x}^-) < 0$. These criteria can be used to predict the binding ability of peptides.

Least square optimization. There are different strategies for the learning phase where the 20*M + 1 free parameters of the scoring function $f(\vec{x})$, eq. (3), are determined. The most elementary approach is to minimize the optimization function with respect to the sum of least squared deviations [least square method (LSM)]

$$L(\vec{w}, b) = \tfrac{1}{2} \sum_{n=1}^{N} \left(f(\vec{x}_n) - y_n \right)^2 . \tag{4}$$

The sum in eq. (4) runs over all sequences of the learning set $\mathbb{S}_{learn} = \mathbb{S}^+ \cup \mathbb{S}^-$, where for binding sequences $y_n = +1$ and for non-binding sequences $y_n = -1$. Taking the derivatives of $L(\vec{w}, b)$ with respect to $\vec{w}$ and b results in the following set of 20*M linear equations

$$\left\langle (\vec{x} - \langle \vec{x} \rangle)\, (\vec{x}^t - \langle \vec{x}^t \rangle) \right\rangle \bullet \vec{w} \;=\; \left\langle \left(y - \langle y \rangle \right) \left(\vec{x} - \langle \vec{x} \rangle \right) \right\rangle \tag{5}$$

and

$$b \;= \langle y \rangle \; - \left\langle \vec{x}^t \right\rangle \bullet \vec{w} \,. \tag{6}$$

The angular brackets in eq. (5) and (6) denote averages over all sequences of the learning set $\mathbb{S}_{learn}$ as for instance

$$\langle \vec{x} \rangle = \tfrac{1}{N} \sum_{n=1}^{N} \vec{x}_n \,. \tag{7}$$

It is interesting to note that the matrix of the set linear equations (5) is formed from the covariances of the sequence distributions

$$\left\langle (\vec{x} - \langle \vec{x} \rangle)\, (\vec{x}^{\,t} - \langle \vec{x}^{\,t} \rangle) \right\rangle = \left\langle \vec{x}\ \vec{x}^{\,t} \right\rangle - \left\langle \vec{x} \right\rangle \left\langle \vec{x}^{\,t} \right\rangle, \tag{8}$$

where $\vec{x}\ \vec{x}^{\,t}$ denotes the dyadic product of the sequence vector $\vec{x}$. For instance the matrix element

$$N \left\langle \vec{x}\ \vec{x}^{\,t} \right\rangle_{(j\,m),\,(j'\,m')}$$

counts how often in the learning set of sequences $\mathbb{S}_{learn}$ one meets an amino acid of type *j* at sequence position *m* while simultaneously at position ***m'*** there is an amino acid of type ***j'***. Hence, the matrix of the set of linear equations (5) accounts for such pair correlations. We have developed our own computer program to solve these linear equations.

Weighting and regularization. To weight binding and non-binding peptides differently one can generalize the averages, eq. (7), according to

$$\langle \vec{x} \rangle = \tfrac{w^+}{N^+} \sum_{n=1}^{N^+} \vec{x}_n^{\,+} + \tfrac{w^-}{N^-} \sum_{n=1}^{N^-} \vec{x}_n^{\,-} \,, \tag{9}$$

where $w^+ + w^- = 1$ and $N^+ + N^- = N$ holds. This description allows a weighting of sequences in the learning set $\mathbb{S}_{learn}$, which is independent from the actual number of binding $\mathbb{S}^+$ and non-binding sequences $\mathbb{S}^-$ considered.

In case the number of data is small compared with the set of parameters that are to be optimized a regularization of the optimization procedure has turned out to be useful. This is the so-called ridge regression procedure (21), which is widely used for sequence prediction problems (17). It can be considered by an additional term in the optimization function, eq. (4), yielding

$$\hat{L}(\vec{w}, b) = L(\vec{w}, b) + \lambda \tfrac{N}{2} \vec{w}^{\,t} \bullet \vec{w} \,, \tag{10}$$

where λ is an empirical parameter, which needs to be chosen. Note that the regularization term is weighted with the total number of sequences N in the learning set $\mathbb{S}_{learn}$ to render the parameter λ independent from the size of the learning set. The regularization term contributes to a minimization of the length of the normal vector $\vec{w}$ of the separating hyperplane that is defined by $f(\vec{x}) = \vec{w}^{\,t} \bullet \vec{x} + b = 0$. As a consequence, the sensitivity of this separating hyperplane may increase for moderate values of λ in particular if the set of linear equations (5) is ill-conditioned due to the smallness of the learning set $\mathbb{S}_{learn}$. Interestingly, a support vector machine uses also a strategy to minimize the hyperplane normal vector $\vec{w}$ to increase the sensitivity (18).

In the set of linear equations the regularization term in the optimization function gives rise to an extra term in the diagonal of the matrix yielding instead of eq. (5)

$$\left\langle (\vec{x} - \langle \vec{x} \rangle)\, (\vec{x}^{\,t} - \langle \vec{x}^{\,t} \rangle) \right\rangle \bullet \vec{w} + \lambda \vec{w} = \left\langle \left(y - \langle y \rangle\right) \left(\vec{x} - \langle \vec{x} \rangle\right) \right\rangle. \tag{11}$$

In the present applications we have a sufficient number of data, such that an application of the ridge regression method did not offer advantages.

Support vector machine. An alternative approach to optimize the parameters of the linear scoring function, eq. (3), is to use a support vector machine (SVM). A detailed description of SVM can be found in Refs. (18, 22). With this method one determines the parameters of the scoring function $f(\vec{x})$ such that for binding peptides the inequality $f(\vec{x}^{\,+}) \geq +1$ and for non-binding peptides the inequality $f(\vec{x}^{\,-}) \leq -1$ holds, while simultaneously the length of the hyperplane normal vector $\vec{w}$ is minimized. The latter increases the sensitivity to discriminate between binding and non-binding peptides. A crucial point of this method is to consider only a subset of the total learning set $\mathbb{S}_{learn}$ by sorting out the data for which the corresponding inequality rigorously holds, i.e. $f(\vec{x}^{\,+}) > +1$ and $f(\vec{x}^{\,-}) < -1$ for binding and non-binding peptides, respectively. Also this selection increases the sensitivity of the method. A further increase in sensitivity may be achieved by applying a non-linear transformation to the feature space of the learning data set and optimizing the discrimination problem in this new feature space. We use the public domain program SVM^{light} (23) to optimize the parameters of the support vector machine.

Quality control. The performance of learning (predicting) can be characterized by providing simply the fraction of binding and non-binding nonapeptides, which were recognized (predicted) properly or alternatively by the Matthew correlation coefficient (MCC) (24), which is defined as

$$\mathrm{MCC} = \frac{\mathrm{cor}^+ \mathrm{cor}^- - \mathrm{incor}^+ \mathrm{incor}^-}{\left[\mathrm{N}^+ \mathrm{N}^- (\mathrm{cor}^+ + \mathrm{incor}^-)(\mathrm{cor}^- + \mathrm{incor}^+)\right]^{\frac{1}{2}}}, \tag{12}$$

with cor^+ and cor^- as correctly classified binding and non-binding peptides and incor^+ and incor^- as incorrectly classified binding and non-binding peptides, respectively. Note that $\mathrm{N}^+ = \mathrm{cor}^+ + \mathrm{incor}^+$ and $\mathrm{N}^- = \mathrm{cor}^- + \mathrm{incor}^-$. The MCC ignores spurious contributions, which are obtained also in the absence of a learning or prediction strategy. In case of a symmetric situation with $\mathrm{cor}^+ = \mathrm{cor}^- = \mathrm{cor}$, $\mathrm{incor}^+ = \mathrm{incor}^- = \mathrm{incor}$, $\mathrm{N}^+ = \mathrm{N}^- = \mathrm{N}/2$ and a low error margin $\mathrm{cor} \gg \mathrm{incor}$ the expression (12) simplifies approximately to

$$\mathrm{MCC} \simeq 1 - 2\frac{\mathrm{incor}}{\mathrm{cor} + \mathrm{incor}}$$

valid for binding and non-binding nonapeptides, such that a prediction probability of 0.9 corresponds to an MCC value of 0.8.

Another widely used method to characterize the quality of learning and predicting are plots of sensitivity (***sens***) versus specificity (***spec***) (25). The functional dependence ***sens(spec)*** can be obtained by varying the threshold t used to classify a peptide of sequence $\bar{\mathrm{x}}$ as binding for $\mathrm{f}(\bar{\mathrm{x}}) > \mathrm{t}$ and as non-binding for $\mathrm{f}(\bar{\mathrm{x}}) < \mathrm{t}$ and monitoring ***sens(t)*** and ***spec(t)***, which are defined as

$$\mathrm{sens}(\mathrm{t}) = \frac{\mathrm{cor}^+(\mathrm{t})}{\mathrm{N}^+} \quad \text{and} \quad \mathrm{spec}(\mathrm{t}) = \frac{\mathrm{cor}^-(\mathrm{t})}{\mathrm{N}^-}. \tag{13}$$

The area under the function ***sens(spec)*** can be understood as an overall quality measure of recognition/ prediction. However, it is preferable to consider the ***sens*** and ***spec*** values for a symmetric situation i.e. ***sens*** ≅ ***spec***, which can be achieved by variation of the threshold t value.

3 Results and Discussion

Parameters of the scoring function. We have determined the parameters of the scoring function by solving the set of linear equations (5) and by applying the support vector machine for different sets of binding and non-binding nonapeptides. To provide useful material for future applications, we calculated the parameters of the scoring function, eq. (3), for all 538 binding nonapeptides $\mathbb{S}^+$ available supplemented by 538 non-binding peptides taken from the set $\mathbb{S}^-$ of 10,000 nonapeptides as described above. The parameters of $\bar{\mathrm{w}}$ and b obtained using the method of minimizing the least square deviation (LSM) and the support vector machine (SVM) are given in Table 2. Surprisingly, the support vector machine did not provide improved results by using a non-linear transformation from which we can conclude that the problem of peptide binding is likely to be not separable in a non-linear feature spaces. Hence, the SVM parameters displayed in Table 2 refer to the linear version.

Learning and recognizing. We discriminate between learning, recognizing and predicting the ability of peptides to bind or not to bind. For the first two procedures we use a learning set $\mathbb{S}_{\mathrm{learn}}$ and for the latter we use a disjoint predicting set $\mathbb{S}_{\mathrm{predict}}$ of binding and non-binding peptides. Determining the parameters given in Table 2, all 538 available binding peptides were used. Hence, in this case we can only probe how well the scoring function recognizes the content of the learning set $\mathbb{S}_{\mathrm{learn}}$ of binding and non-binding nonapeptides. With LSM the 538 binding peptides were recognized with a probability of 0.916 and the 538 non-binding peptides with probability of 0.952. For the SVM the corresponding probabilities are 0.931 and 0.925. The support vector machine was using 197 binding and 202 non-binding nonapeptides yielding a total number of 399 support vectors. The absolute numbers of incorrectly as non-binding recognized truly binding peptides are 37 and 46 for SVM and LSM, respectively. From these, 32 peptides were incorrectly recognized with both methods. The absolute numbers of incorrectly recognized peptides from the truly non-binding peptides are 40 for SVM and 24 for LSM. In this case only 4 peptides were recognized wrongly as binding peptides with both methods.

Table 2. Optimized parameters $\bar{w}$ of the scoring function $f(\bar{x}) = \bar{w}^t \bullet \bar{x} + b$.

amino acid	position 1		position 2		position 3		position 4		position 5		position 6		position 7		position 8		position 9	
	LSD	SVM	LSD	SVM	LSD	SVM	LSD	SVM	LSD	SVM	LSD	SVM	LSD	SVM	LSD	SVM	LSD	SVM
ALA	0.15	0.08	0.28	0.00	0.30	0.23	0.05	0.03	0.06	-0.07	0.16	0.16	0.18	0.02	0.15	-0.05	-1.70	0.18
HIS	-0.07	0.12	-0.01	-0.02	0.22	0.00	0.12	-0.22	0.24	-0.20	0.31	0.19	0.09	-0.04	0.14	-0.18	-1.84	-0.29
GLU	-0.15	-0.24	0.30	-0.32	0.03	-0.44	0.20	0.03	0.16	-0.17	0.13	-0.16	-0.10	-0.28	0.12	0.06	-1.74	-0.16
GLN	0.09	-0.13	0.19	-0.25	0.12	-0.06	0.20	0.23	0.12	0.11	0.04	-0.13	0.02	0.12	0.14	0.04	-1.70	0.15
ASP	0.00	-0.29	0.13	-0.32	0.30	-0.03	0.29	0.27	-0.07	-0.18	-0.02	-0.21	-0.08	-0.34	0.15	-0.07	-1.97	-0.32
ASN	-0.19	-0.21	-0.14	-0.32	0.33	0.19	-0.06	-0.13	0.05	0.10	0.19	-0.11	-0.08	-0.18	-0.09	-0.04	-1.98	-0.35
LEU	0.03	-0.09	1.22	1.76	0.33	0.16	-0.11	-0.05	0.05	0.12	0.32	0.07	0.19	0.39	0.21	0.38	-1.35	0.55
GLY	0.08	0.02	0.05	-0.58	0.19	0.17	0.26	0.27	0.33	0.36	0.21	-0.07	0.08	-0.31	0.09	-0.02	-1.92	-0.34
LYS	0.19	0.15	0.07	-0.22	0.34	-0.22	0.29	0.21	-0.06	-0.12	0.03	-0.32	-0.31	-0.31	0.10	0.05	-1.84	-0.24
SER	0.12	0.18	0.03	-0.51	0.33	-0.25	0.21	0.12	-0.01	-0.18	0.24	0.07	0.03	0.05	0.20	0.15	-1.75	-0.38
VAL	0.23	0.03	0.48	0.34	0.18	-0.13	0.17	-0.15	0.33	0.29	0.34	0.24	0.12	0.28	0.04	-0.11	-1.15	0.93
ARG	0.28	0.09	-0.01	-0.21	0.26	-0.12	0.25	0.11	0.14	0.01	0.07	-0.09	-0.03	-0.33	0.08	0.20	-1.86	-0.05
THR	0.06	-0.12	0.42	0.04	0.12	-0.19	0.12	-0.25	0.10	-0.01	0.11	-0.04	0.04	-0.25	0.33	0.00	-1.63	0.17
PRO	-0.06	0.00	-0.01	-0.15	0.00	-0.13	0.29	0.22	0.16	0.03	0.10	0.16	0.24	0.26	0.20	0.01	-1.93	-0.24
ILE	0.08	0.10	0.68	0.63	0.15	0.10	0.25	0.07	0.26	0.06	0.39	0.43	0.21	0.32	0.20	0.17	-1.42	0.32
MET	0.21	0.09	1.15	1.04	0.27	0.22	-0.13	-0.11	-0.16	-0.04	0.03	0.10	0.01	-0.11	0.11	-0.12	-1.53	0.02
PHE	0.22	0.42	0.20	-0.39	0.37	-0.03	-0.07	-0.26	0.19	0.04	0.32	0.00	0.32	0.40	0.37	0.03	-1.70	0.16
TYR	0.21	-0.06	0.20	-0.11	0.23	0.07	0.03	0.02	0.08	0.10	0.09	-0.12	0.14	-0.20	0.09	0.00	-1.89	-0.11
CYS	0.04	-0.19	0.48	-0.23	0.11	0.08	0.00	-0.17	0.08	-0.09	0.22	-0.17	0.15	0.11	0.30	-0.31	-1.56	0.16
TRP	-0.17	0.08	0.05	-0.19	0.58	0.38	0.15	-0.23	0.28	-0.17	0.13	0.00	0.04	0.37	0.06	-0.19	-1.89	-0.15

The 180 parameters are displayed in a two-dimensional array w_{jm} with respect to the 20 amino acid types ***j*** and the 9 sequence positions ***m***. The parameters obtained with the least square optimization method (LSM) are given in the left columns the parameters from the support vector machine (SVM) are given in the right columns, respectively. The values of b are $b_{LSD} = -0.31$ and $b_{SVM} = 1.27$. The weights used to compute the averages, eq. (9), needed for LSM were $w^+ = 0.36$ and $w = 0.64$ for the binding and non-binding peptides, respectively.

Table 3. Number of incorrectly recognized sequences and the corresponding range of values of the scoring function for different weights.

weights w^+	incorrect binding	scoring function	incorrect non-bind.	scoring function
0.8	**0**	**–**	**17**	**0.01 to 0.98**
0.6	**1**	**at –0.02**	**7**	**0.02 to 0.72**
0.4	**7**	**–0.36 to –0.07**	**2**	**at 0.23, at 0.45**
0.2	**21**	**–0.77 to –0.02**	**1**	**at 0.11**
0.1	**37**	**–1.02 to –0.02**	**0**	**–**

Weighting binding and non-binding peptides. The least square optimization method allows to apply different weights for the set of binding and non-binding peptides [see eq. (9)]. These weights can play a similar role as does the threshold value t [see eq. (13)] used to discriminate between binding and non-binding peptides. We study the influence of different weights on recognition by monitoring the scoring function $f(\bar{x})$, eq. (3), for renumbered sequences $\bar{x}_n$ of the learning set $\mathbb{S}_{learn}$, which are ordered such that for subsequent sequences $\bar{x}_n$ and $\bar{x}_{n+1}$ we have $f(\bar{x}_n) < f(\bar{x}_{n+1})$. Thus, a scoring function is obtained whose value increases monotonously with sequence number n. The scoring function shown in Figure 1 is based on 200 binding and 200 non-binding peptides in the learning set $\mathbb{S}_{learn}$ to determine the parameters of the scoring function and its values. The crosses mark sequences whose peptides were recognized incorrectly. The number of incorrect recognized binding (non-binding) sequences increases from 0 to 37 (decreases from 17 to 0) with decreasing weight w^+ for the binding peptides (Table 3). The ideal shape of the scoring function should be a step function with a function value of –1 for the first 200 non-binding peptides and +1 for the 200 binding peptides. For large weights w^+ of the binding peptides the positive step of the scoring function is very pronounced while the negative step is less distinct. The opposite is the case for small weights w^+.

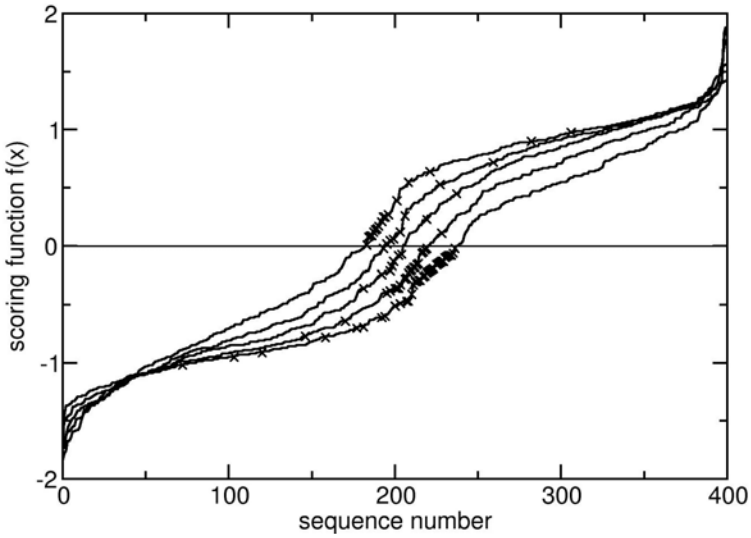

Figure 1. Course of the scoring function $f(\vec{x})$, eq. (3), for different weights w^+ of the binding peptides. Parameters of the scoring function were determined based on 200 binding and 200 non-binding peptides in the learning mode as explain in text. The scoring function is displayed in recognition mode considering the peptides of the learning set $\mathbb{S}_{learn}$. From top to bottom the scoring functions refer to weights w^+ of the binding peptides of 0.8, 0.6, 0.4, 0.2, 0.1. Crosses mark incorrectly predicted peptides whose statistics are given in Table 3.

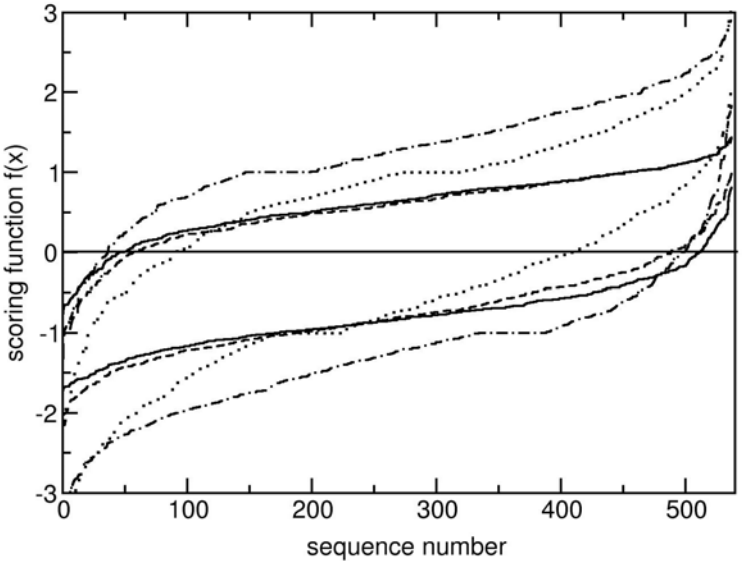

Figure 2. Course of the scoring function $f(\vec{x})$, eq. (3), monitored separately for binding and non-binding peptides. In contrast to Figure 1 binding and non-binding peptides were renumbered separately and the resulting scoring functions were displayed separately. From the pair of scoring functions displayed in the same line style the lower $f^-(\vec{x})$ refers to non-binding peptides and the upper $f^+(\vec{x})$ to binding peptides. For the learning set all available 538 binding peptides and the same number of non-binding peptides chosen from the set of 10,000 random peptides were considered. For LSM . The dashed-dotted lines display results of recognition using SVM. The pair of solid lines describe the same as before but using least square optimization (LSM), which is also used for all other data displayed in this figure. The pair of dashed lines displayed results in prediction mode obtained with the jackknife method. The dotted lines display the prediction of peptides binding using only 50 binding and 50 non-binding peptides.

Course of the scoring function. To study the behavior of the scoring function in more detail we employed a learning set of all available 538 binding peptides and added the same number of non-binding peptides. As in Figure 1 we renumbered the sequences to obtain monotonously increasing scoring functions. But, in this case we considered the binding and non-binding peptides separately yielding two branches $f^+(\vec{x})$ and $f^-(\vec{x})$ of the scoring function, respectively. The branches $f^-(\vec{x})$ of the non-binding peptides are located in the lower half, the branches $f^+(\vec{x})$ describing the binding peptides are located in the upper half of Figure 2. The fraction

of non-binding peptides with $f^-(\bar{x}) < 0$ and of binding peptides with $f^+(\bar{x}) > 0$ are correctly recognized/predicted. In recognition mode (solid lines for LSM and dashed-dotted lines for SVM) the two different optimization methods (LSM and SVM) considered in this work yielded very similar results with a minor advantage for SVM in recognizing binding peptides, while LSM is marginally ahead in recognizing non-binding peptides. But, SVM seems to be superior in its ability to separate binding from non-binding peptides, since its scoring function is generally larger for binding peptides and smaller for non-binding peptides as compared to the corresponding LSM scoring function. The results obtained with LSM in prediction mode using the jackknife procedure (dashed lines) (leaving out one peptide in the learning mode, whose binding ability is predicted) yielded results that are very similar to the corresponding data obtained in recognition mode. Even with a rather small number of 50 binding and 50 non-binding peptides in the learning set, prediction of all 538 binding and non-binding peptides yields reasonable results (dotted lines).

Selecting peptides from the learning set. The support vector machine has the ability to select a subset of data in feature space to optimize the performance. The least square optimization method does not directly offer such an option. However, after an LSM optimization is performed one can identify incorrectly recognized peptides and the peptides that are located in the twilight zone of vanishing values of the scoring function. The assumed binding ability of these peptides may have been wrongly assigned. This can particularly be the case for the randomly generated sequences from which we assumed that they are all non-binding. We have the option to eliminate these peptides in a second run of LSM optimization. To investigate this possibility, we started an LSM optimization with 300 binding and 5,000 non-binding peptides as learning set $\mathbb{S}_{learn}$ with $w^+ = 0.36$. In the prediction mode, we considered the remaining 238 binding and 5,000 non-binding peptides. Thus, in recognition mode 92.0% binding and 92.8% non-binding peptides were found. In the prediction mode 90.3% binding and 92.5% non-binding peptides were predicted correctly. In a second LSM run, we eliminated 31 non-binding peptides with $f(\bar{x}^-) > 0.7$ from $\mathbb{S}_{learn}$. With this choice, recognition is slightly reduced for the non-binding peptides yielding 92.5%, while it is unchanged for the binding peptides. In the prediction mode we now obtain that 91.6% of the binding and 92.2% of the non-binding peptides are predicted correctly, which is over all an improvement compared with the first LSM run.

Interestingly, the SVM optimization yielded only 50.6% correctly recognized and 48.7% correctly predicted binding peptides, while the non-binding peptides were found with 99.5% in recognition and prediction mode. The reason for this failure of SVM is the uneven size of the set of binding and non-binding peptides in this application. We can artificially increase the set of binding peptides in the SVM procedure by considering 16 identical copies of the original set of 300 binding peptides to symmetrize the number of binding and non-binding peptides considered. Astonishingly, in this case we obtain correct recognition of 96.0% binding and 92.0% non-binding peptides and correct prediction of 93.0% binding and 91.0% non-binding peptides. In the LSM optimization a balanced consideration of binding and non-binding peptides in the learning mode can be achieved directly by the weights w^+ and w^-, which are used to evaluate the averages, eq. (9). In conclusion one can say that with a well tuned least square optimization the same quality of predicting of peptides can be achieved as with the SVM optimization.

Quality control. A sensitivity selectivity plot (Figure 3) can be used as quality control for recognition and prediction. The area under the function sensitivity(specificity) provides an overall measure of quality. The area is 0.992 for recognition (dashed line) and 0.938 for prediction (solid line) using the LSM optimization and 0.965 for prediction (dotted line) using SVM optimization. Here, SVM shows again its superiority being stronger in its ability to discriminate clearly between binding and non-binding peptides.

To obtain a more reliable measure of prediction quality, we considered randomly chosen sets of peptides for learning and predicting considering the LSM optimization. To obtain good statistics we determined the parameters of the scoring function 400 times with different learning sets randomly chosen from the total number of 538 binding peptides and 10,000 non-binding peptides. Table 4 shows the results for learning sets of 50 binding and 50 non-binding peptides, which yield perfect results for recognition, since the set is so small, but exhibit rather modest results in the prediction mode with an average success below 80% and a large variance. With a much larger learning set of 400 binding and 5,000 non-binding peptides the average fraction of correctly recognized peptides diminishes being now above 90%, but at the same time the average prediction improves considerably being now close to 90% for binding and non-binding peptides. The variance of these averages is now much smaller due to the larger data base of binding and non-binding peptides used for the determination of the parameters of the scoring function. Due to these variances the actual uncertainty of a prediction may be larger than the average fraction of correct predictions.

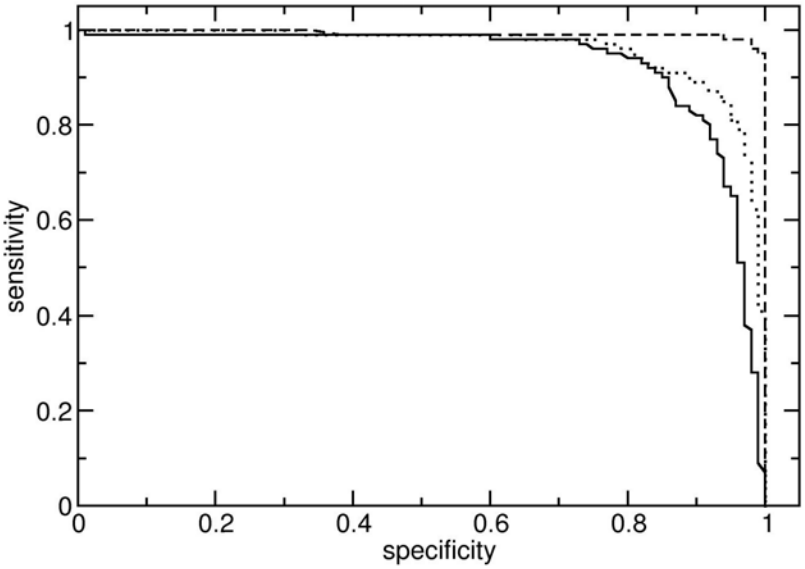

Figure 3. Sensitivity specificity plot for two disjoint sets of 200 binding and 200 non-binding peptides for learning and predicting mode, respectively. Dashed line: learning mode, solid line: predicting mode with LSM optimisation. Dotted line: SVM in predicting mode.

Table 4. Recognition and prediction statistics of binding for different learning sets of peptides [a]

	size of learning sets binding/non-binding [b] 50/50		size of learning sets binding/non-binding [b] 400/5,000	
	binding peptides	non-binding peptides	binding peptides	non-binding peptides
recognition [c]	100% ±0.0%	100% ±0.0%	92.0% ±0.6%	92.8% ±0.1%
prediction [d]	78.3% ±26.5%	72.8 ±19.8%	88.5% ±6.5%	92.4% ±0.2%

[a] **The learning sets are generated at random 400 times using least square optimization (LSM).**
[b] **Number of binding and non-binding peptides.**
[c] **Recognition mode: probing recognition probability of the different learning sets of peptides.**
[d] **Prediction mode: probing prediction probability of randomly chosen 138 binding and 5,000 non- binding peptides that are disjoint from the learning set.**

4 Conclusions

We have generalized a least square optimization method to predict peptide binding at the class I major histocompatibility complex using a general linear scoring function. A new weighting procedure allows to treat asymmetric data sets with a small number of binding and a large number of non-binding peptides in a balanced way maintaining the prediction quality, which is in the case far better than the results from the support vector machine (SVM). However, the apparent deficiency of SVM can probably be repaired by generalizing existing programs solving the SVM problem. But, even for a symmetric data set the prediction quality of LSM comes very close to the SVM results. Further generalizations of the LSM may possess the potential to reach or surpass the prediction quality of SVM.

Acknowledgements. We are grateful for financial support from the Deutsche Forschungsgemeinschaft Sfb498, GRK80/2, GRK268, GRK788/1, Forschergruppe 475.

References

1. Guilloux, Y., Lucas, S., Brichard, V. G., Van Pel, A., Viret, C., De Plaen, E., Brasseur, F., Lethe, B., Jotereau, F., and Boon, T. A peptide recognized by human cytolytic T lymphocytes on HLA-A2 melanomas is encoded by an intron sequence of the N-acetylglucosaminyltransferase V gene. *J Exp*

Med, 183:1173-1183, 1996.
2. Garboczi, D. N., Ghosh, P., Utz, U., Fan, Q.R., Biddison, W.E., and Wiley, D.C. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature*, 384:134-141, 1996.
3. Lanzavecchia, A., Reid, P.A., and Watts, C. Irreversible association of peptides with class II MHC molecules in living cells. *Nature*, 357:249-252, 1992.
4. Stolze, L., Nussbaum, A.K., Sijts, A., Emmerich, N.P., Kloetzel, P.M., and Schild, H. The function of the proteasome system in MHC class I antigen processing. *Immunol. Today*, 21:317-319, 2000.
5. Garcia, K. C., Degano, M., Pease, L.R., Huang, M., Peterson, P.A., Leyton, L., and Wilson, I.A. Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. *Science*, 279:1166-1172, 1998.
6. Tortorella, D., Gewurz, B.E., Furman, M.H., Schust, D.J., and Ploegh, H.L. Viral subversion of the immune system. *Annu Rev Immunol*, 18:861-926, 2000.
7. Bouvier, M., and Wiley, D.C. Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules. *Science*, 265:398-402, 1994.
8. Falk, K., Rötzschke, O., Stefanovic, S., Jung, G., and Rammensee, H.G. Allele-specific motifs revealed by sequencing of self peptides eluted from MHC molecules. *Nature*, 351:290-296, 1991.
9. Regner, M., Claesson, M.H., Bregenholt, S., and Röpke, M. An improved method for the detection of peptide-induced upregulation of HLA-A2 Molecules on TAP-deficient T2 cells. *Exp Clin Immunogenet*, 13:30-35, 1996.
10. Henderson, R. A., Michel, H., Sakaguchi, K., Shabanowitz, J., Apella, E., Hunt, D.F., and Engelhard, V.H. HLA-A2.1 associated peptides from a mutant cell line. A second pathway of antigen presentation. *Science*, 255:1264-1266, 1992.
11. Rammensee, H. G., Bachmann, J., Emmerich, N.N., Bachor, O.A., and Stevanovic, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50:213-219, 1999.
12. Brusic, V., Rudy, G., and Harrsison, L.C. MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Research*, 26:368-371, 1998.
13. Rosenfeld, R., Zheng, Q., Vajda, S., and Delisi, C. Computing the structure of bound peptides - application to antigen recognition by class-I major histocompatibility complex receptors. *Journal of Molecular Biology*, 234:515-521, 1994.
14. Rotzschke, O., Falk, K., Stevanovic, S., Jung, G., and Rammensee, H.G. Peptide motifs of closely related HLA class-I molecules encompass substantial differences. *European Journal of Immunology*, 22:2453-2456, 1992.
15. Gulukota, K., Sidney, J., Sette, A., and DeLisi, C. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *Journal of Molecular Biology*, 267:1258-1267, 1997.
16. Brusic, V., Rudy, G., Honeyman, M., Hammer, J., and Harrison, L. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics*, 14:121-130, 1998.
17. Peters, B., Bulik, S., Tampe, R., van Endert, P.M., and Holzhutter, H.G. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors´. *Journal of Immunology*, 171:1741-1749, 2003.
18. Hearst, M. A., Schölkopf, B., Dumais, S., Osuna, E. and Platt, J. Trends and Controversies - Support Vector Machines. *IEEE Intelligent Systems*, 13:18-28, 1998.
19. Dönnes, P., and Elofsson, A. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics 3*, 3, 2002.
20. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne P.E. The Protein Data Bank. *Nucleic Acids Research*, 28:235-242, 2000.
21. Hoerl, A. E. a. K., R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55-67, 1970.
22. Burges, C. J. C. A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*121-167, 1998.
23. Joachims, T. Making large-Scale SVM Learning Practical. *In:* B. S. a. C. B. a. A. Smola (ed.), Advances in Kernel Methods - Support Vector Learning.: MIT-Press, 1999.
24. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.*, 405:442-451, 1975.
25. Bradley, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145-1159, 1997.

doi:10.1016/j.jmb.2006.01.064 J. Mol. Biol. (2006) 357, 1511–1521

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Solution Structure and Backbone Dynamics of the *Trypanosoma cruzi* Cysteine Protease Inhibitor Chagasin

Didier Salmon[1], Rodolpho do Aido-Machado[1], Anne Diehl[2] Martina Leidert[2], Oliver Schmetzer[3], Ana P. C. de A. Lima[4] Julio Scharfstein[4], Hartmut Oschkinat[2] and José R. Pires[1]*

[1]*Instituto de Bioquímica Médica CCS, Universidade Federal do Rio de Janeiro, Av. Brigadeiro Trompowiski s/n, Rio de Janeiro RJ 21941-590, Brazil*

[2]*Forschungsinstitut für molekulare Pharmakologie Robert-Rössle-Str. 10, 13125 Berlin, Germany*

[3]*Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany*

[4]*Instituto de Biofísica Carlos Chagas Filho, CCS, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ 21949-900, Brazil*

**Corresponding author*

A *Trypanosoma cruzi* cysteine protease inhibitor, termed chagasin, is the first characterized member of a new family of tight-binding cysteine protease inhibitors identified in several lower eukaryotes and prokaryotes but not present in mammals. In the protozoan parasite *T. cruzi*, chagasin plays a role in parasite differentiation and in mammalian host cell invasion, due to its ability to modulate the endogenous activity of cruzipain, a lysosomal-like cysteine protease. In the present work, we determined the solution structure of chagasin and studied its backbone dynamics by NMR techniques. Structured as a single immunoglobulin-like domain in solution, chagasin exerts its inhibitory activity on cruzipain through conserved residues placed in three loops in the same side of the structure. One of these three loops, L4, predicted to be of variable length among chagasin homologues, is flexible in solution as determined by measurements of ^{15}N relaxation. The biological implications of structural homology between chagasin and other members of the immunoglobulin super-family are discussed.

Keywords: chagasin; *T. cruzi*; cysteine protease inhibitors; NMR structure; dynamics

Introduction

The unicellular protozoan pathogen *Trypanosoma cruzi* is the causative agent of Chagas disease, a human parasitic illness endemic in Central and South America that affects 16–18 million people while leaving over 100 million at risk†. To date, no vaccines are available and drugs for the treatment are inadequate. New perspectives for drug treatment of Chagas disease came from analysis of the structure and function of the major lysosomal *T. cruzi* cysteine protease, cruzipain.[1–4] The recent sequencing of the *T. cruzi* genome revealed that Clan CA cysteine proteases (CP) belonging to the C1 papain-like family are well represented in the parasite genome.[5] Accordingly, members of the C1 CP family in *T. cruzi* include a single copy gene encoding a cathepsin B-like protease and at least 11 polymorphic genes encoding the cathepsin L-like cruzipain. Synthesized as zymogens, pro-cruzipain is converted into active enzyme following proteolytic cleavage of the N-terminal pro-peptide domain. The maturation process, initiated during trafficking through the Golgi[6] is completed by delivery of cruzipain into the lysosomal compartment.[7] Cruzipain is expressed in all parasite life stages and responsible for the major proteolytic activity in the insect-stage of the parasite.[8,9] Multiple lines of evidence indicated that cruzipain function is crucial for parasite infectivity and survival in mammalian host cells.[1,8,10,11]

Abbreviations used: CP, cysteine protease; ICP, inhibitor of cysteine peptidase; Ig, immunoglobulin; HSQC, heteronuclear single quantum coherence; NOE, nuclear Overhauser enhancement; NOESY, NOE spectroscopy; CSP, chemical shift perturbation; TOCSY, total correlation spectroscopy.

E-mail address of the corresponding author: jrmpires@gmx.net

† http://www.who.int/ctd/chagas

0022-2836/$ - see front matter

The determination of the X-ray structure of the catalytic domain of cruzipain[2] bound to an irreversible inhibitor paved the way for the development of non-toxic drug analogues, some of which proved capable of protecting mice from lethal *T. cruzi* infections.[3]

To date, 48 families of protease inhibitors are described, containing not less than 200 members in mammalian cells (e.g. serpins and cystatins).[12] In trypanosomatids, genes coding for CP inhibitors homologous to the mammalian cystatins are absent.[13] Instead, trypanosomatids, amoeba[14] and some prokaryotes[15,16] express CP inhibitor proteins that share between 20% and 40% sequence identity with chagasin,[15] a single-chained 110 amino acid residue protein originally identified in *T. cruzi*. This family is termed as ICP (inhibitors of cysteine peptidases).[17] Studies of the functional role of chagasin demonstrated that it modulates the endogenous activity of cruzipain, thus indirectly interfering with *T. cruzi* ability to differentiate and/or to invade mammalian host cells.[18] Along similar lines, recent data suggest that chagasin homologues in *Leishmania mexicana* modulate the outcome of host–parasite interactions.[19]

Threading and comparative modelling provided predictions of the structure of chagasin-like proteins.[20] Based on these theoretical studies, it was hypothesized that chagasin-like ICPs adopt an immunoglobulin-like (Ig-like) fold. It was further proposed that some conserved residues in loop regions (L2, L4 and L6; Figure 1) of chagasin could be implicated in the inhibitory interaction with CPs.[15] Nevertheless, due to the low level of sequence identity of chagasin with the templates used in the abovementioned modelling studies (ca 12–17%), it was important to validate these predictions by direct structural data. Here, we determined the solution structure of *T. cruzi* chagasin, studied its backbone dynamics and mapped its interaction with cruzipain by ^{15}N-heteronuclear single quantum coherence (HSQC) based experiments.

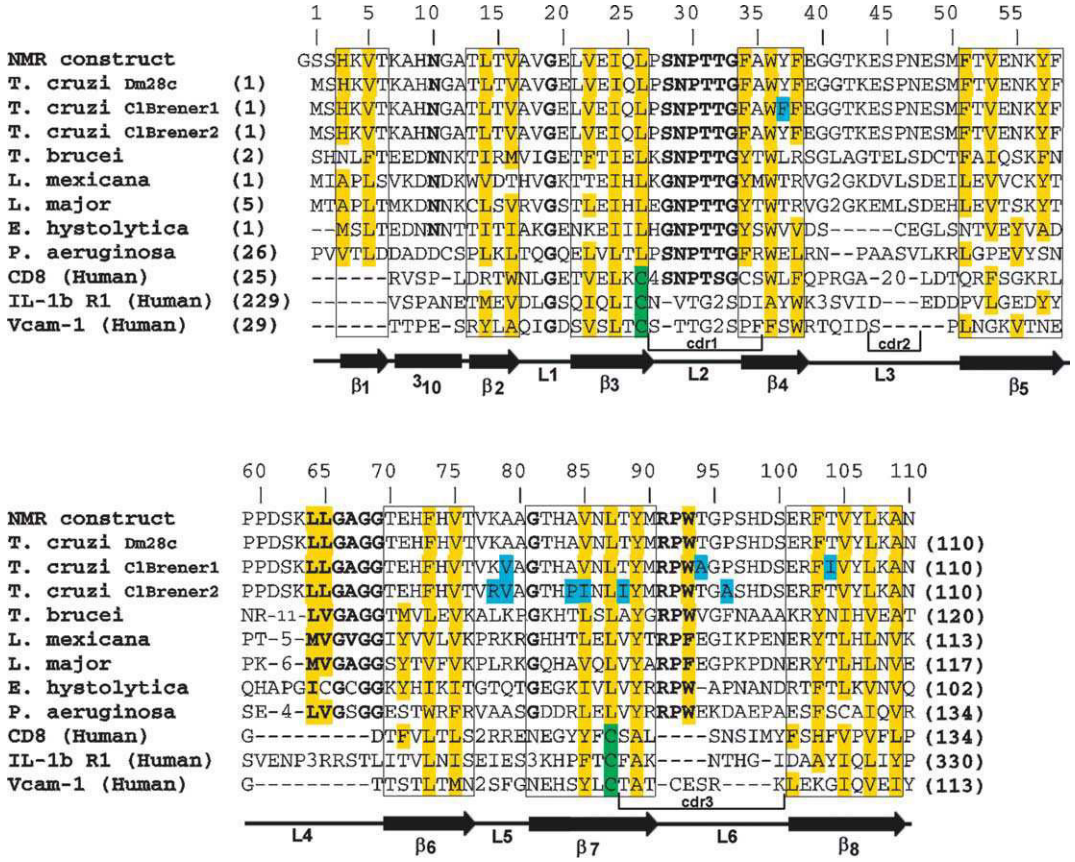

Figure 1. Sequence alignment of chagasin-like cysteine protease inhibitors from different pathogen bacteria and protozoa and of structurally homologous human proteins. Residue numbers refer to *T. cruzi* chagasin. β-Strands and 3_{10}-helix are displayed by arrows and cylinders, respectively, as found in the *T. cruzi* chagasin NMR structure, and the sequences relative to the β-strands are boxed. The positions of the most conserved residues in loops are highlighted in bold and conserved hydrophobic (aromatic and aliphatic) residues are coloured yellow. *T. cruzi* non-conserved residues are highlighted in blue. Complementarity-determining regions (CDR 1–3) in CD8 are indicated. Swiss-Prot accession codes are as follows: chagasin-like ICPs from *T. cruzi* (Dm28c clone, Q966X9; CL Brenner strain, isoform 1, Q4DH32; CL Brenner strain, isoform 2, Q4DY71); *T. brucei* (Q868H0), *L. mexicana* (Q868H1), *L. major* (Q868G9), *E. hystolytica* (Q6KCA4), *P. aeruginosa* (Q9I5G0) and human structural homologue proteins CD8 T-cell surface glycoprotein (P01732), IL-1 R1, interleukine-1 type1 receptor (P14778); Vcam-1, vascular cell adhesion molecule-1 (P19320).

Results

NMR structure of chagasin from *T. cruzi*

The sequence of the chagasin construct used in our NMR experiments is shown in Figure 1. The solution structure of chagasin was determined by multidimensional NMR spectroscopy[21,22] on 0.52 mM ^{15}N-labelled, 0.73 or 0.3 mM ^{13}C, ^{15}N-labelled and 1.19 mM unlabelled protein samples. The assignments of distance restraints derived from NOEs were made in a semi-automated fashion. Initially, we manually assigned 36 distances from NOEs characteristic of secondary structure involving amide–amide, amide–H^{α} or H^{α}–H^{α} protons. Additionally, 39 hydrogen bonds were assigned from slow-exchanging amide protons identified in a 2D ^{15}N HSQC spectrum following exchange into $^{2}H_2O$. Figure 2 summarizes schematically these initial manual assignments and the resulting secondary topology of chagasin. Further nuclear Overhauser enhancements (NOEs) were manually assigned involving some side-chain hydrogen atoms with very well resolved resonances (e.g. I24, L26, L87 methyl groups and Y89 and R91 methylene groups). In total, ca 150 distance restraints were manually assigned. Nevertheless, the great majority of NOE restraints were assigned automatically using the program ARIA,[23,24] yielding in total 2692 restraints. A complete count of restraints used in the final structure calculation is listed in Table 1.

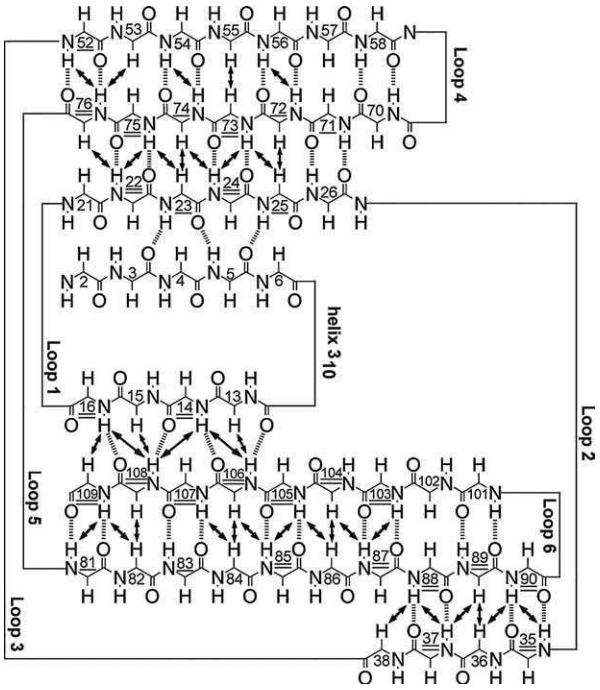

Figure 2. Secondary structure of chagasin. The backbone of residues in β-strands are represented and numbered (β1 3–6, β2 13–16, β3 21–26, β4 35–38, β5 52–58, β6 70–76, β7 81–90 and β8 101–109). Manually assigned NOEs are shown by arrows, hydrogen bonds assigned on the basis of H/^{2}H exchange experiments are shown by dotted lines. Residues underlined have shown H/^{2}H exchange rates of the order of hours while double-underlined residues have shown exchange rates of the order of days.

The ensemble consisting of the 15 lowest-energy structures is shown in Figure 3(a) and a ribbon representation of the lowest-energy structure is shown in Figure 3(b). This ensemble has a mean backbone root-mean-square deviation (rmsd) to the average structure of 0.40 Å for the well-structured regions (residues 3–58 and 70–109). The backbone rmsd over the residues 3–109 is 1.25 Å. Chagasin's tertiary structure is well defined by an average of 9.5 medium or long-range NOE distance restraints per residue. Detailed structure statistics are shown in Table 1.

Chagasin consists of eight β-strands and one small 3_{10} helix. The β1 comprises residues H3 to T6, the helix 3_{10} residues K7 to A12, β2 extends from residues T13 to V16, β3 from residues L21 to L26, β4 from residues A35 to F38, β5 from residues T52 to F58, β6 from residues T70 to T76, β7 from residues G81 to M90 and β8 from residues E101 to A109. The three-dimensional arrangement of the strands is similar to that of an Ig-like domain. The strands are organized into two β-sheets containing four strands each and forming a "Greek-key" β-sandwich. In the first β-sheet, β1 contacts β3 in parallel orientation, β3 is additionally antiparallel to β6, which is antiparallel to β5. In the second β-sheet, β2 is

Table 1. Structure statistics

A. *Number of restraints*	
NOE restraints	2692
Intra-residual	763
Sequential	481
Medium-range	215
Long-range	836
Ambiguous	397
Hydrogen bonds (two restraints each)	39
Dihedral angles (ϕ, ψ)	106
Total	2876
B. *rmsd from experimental restraints*	
NOEs (Å)	0.05±0.02
Dihedral angles (°)	1.5±0.4
C. *CNS potential energy (kcal mol^{-1})*	
E_{total}	−2912±125
E_{bonds}	51±6
E_{angles}	242±15
$E_{impropers}$	163±18
$E_{dihedral}$	557±7
E_{vdW}	−235±29
E_{elec}	−3900±87
E_{noe}	195±48
E_{cdih}	15±11
D. *rmsd (Å) between average structure and the ensemble*[a]	
Backbone	0.40±0.07
All non-H	0.79±0.12
E. *Ramachandran plot analysis* (%)[a]	
Residues in most favored regions	76.3
Residues in additionally allowed regions	19.8
Residues in generously allowed regions	2.1
Residues in disallowed regions	1.9

[a] Excluding flexible loop L4 (residues 59–69).

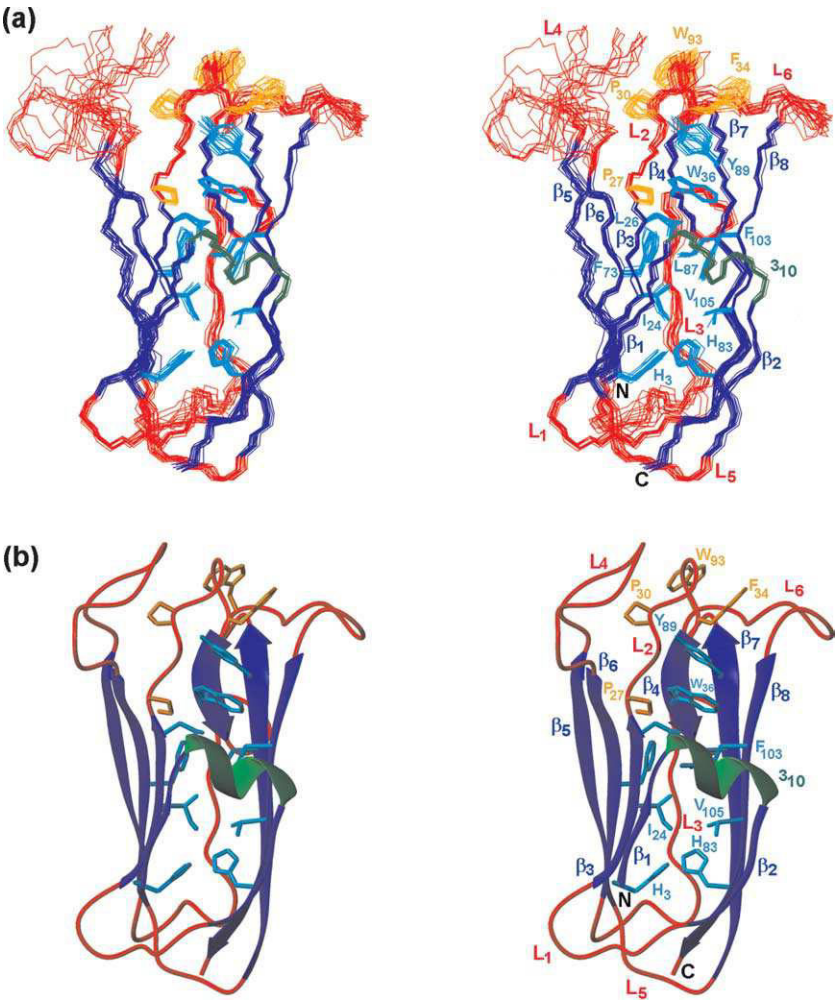

Figure 3. Solution NMR structure of the *T. cruzi* chagasin (stereo view). (a) Superposition of the backbone atoms for the 15 lowest-energy structures of chagasin, residues 3–110. β-Strands (β1–β8), 3_{10}-helix, and loops (L1–L6) are coloured blue, green and red, respectively. Selected residue side-chains are displayed and labelled. The structural statistics are given in Table 1. (b) Ribbon diagram of the chagasin lowest-energy structure. The colouring scheme is the same as in (a).

oriented parallel with β8, which is additionally antiparallel to β7, which is in turn antiparallel to β4. The two β-sheet planes are nearly antiparallel to each other. The helix 3_{10} and four of the six loops (L1, L2, L3 and L5) are crossing from one β-sheet to the other.

The loop regions are only slightly less convergent than the β strands or the helix 3_{10}, except loop L4 (residues 59–69), which was poorly defined. Figure 4 shows the number of NOEs observed as well as the average local rmsd as a function of residue number. Loop 4 is characterized by only very few medium and long-range NOEs (Figure 4(a)) and higher local average rmsd values were obtained in this region for the corresponding NMR structures (Figure 4(b)).

Dynamics of chagasin from NMR measurements

To assess the degree of internal mobility of chagasin, we performed ^{15}N T_1 and T_2 relaxation and ^{1}H–^{15}N heteronuclear NOE experiments (Figure 5). These data showed that residues at the N terminus (residues 1–3) were flexible. In addition, the relaxation data showed decreased ^{15}N R_2 values (Figure 5(b)) and more negative ^{1}H–^{15}N heteronuclear NOEs (Figure 5(a)) for residues in loop 4 (residues 59–69), indicating internal mobility of this region.

A quantitative analysis for global and internal motions was performed using the Lipari & Szabo model free dynamics formalism[25] as implemented in the program Tensor2.[26] At 0.52 mM average values of relaxation rates ($R_1 = 0.88\ s^{-1}$ and $R_2 = 15.9\ s^{-1}$) were calculated considering all the residues in secondary structure elements and from its ratio ($R_2/R_1 = 18.0$) the global correlation time for isotropic tumbling were estimated ($\tau_c = 13$ ns). S^2, the square of the order parameter, was typically between 0.8 and 0.9 for most residues, including loop regions, except for residues in loop 4 (e.g. L64 to G68) where an S^2 value between 0.65 and 0.75 and an effective correlation time (τ_e) around 50 ps were necessary to fit the relaxation data. Additionally for residues at the N terminus (Ser1 and Ser2) the values 0.51 and 0.55 were obtained for S^2, respectively. These results were unaffected by the model considered, either isotropic or anisotropic tumbling. Thus, the differential relaxation behavior of loop 4 and the N-terminal residues are not related to contributions of anisotropic motion and these regions are likely to be genuinely flexible in solution.

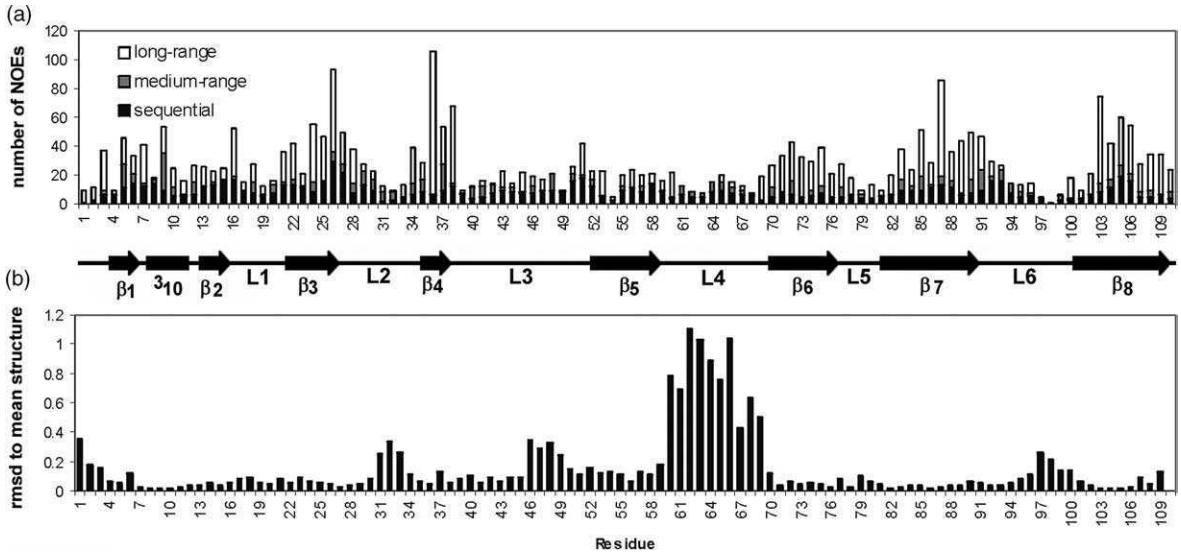

Figure 4. Precision of the chagasin structure. (a) Number of sequential, medium-range, and long-range NOE distance restraints per residue. β-Strands, 3_{10}-helix and loops are indicated schematically. (b) Local rmsd values of backbone atoms for residues 1–110 of chagasin. The 15 lowest-energy structures are considered.

The signals in the NOESY spectra from residues in loop 4 were consistently weak or absent, indicating weak dipolar interactions and consequently poor cross-relaxation of nuclei in loop 4 with neighbouring regions. Consequently, this region of the molecule is very poorly restrained in the NMR structure (Figure 4(a)).

Mapping of the interaction between chagasin and cruzipain

In an attempt to map the chagasin regions interacting with the *T. cruzi* cysteine protease cruzipain, three samples were prepared containing, respectively, 100 μM ^{15}N-labelled chagasin and 22,

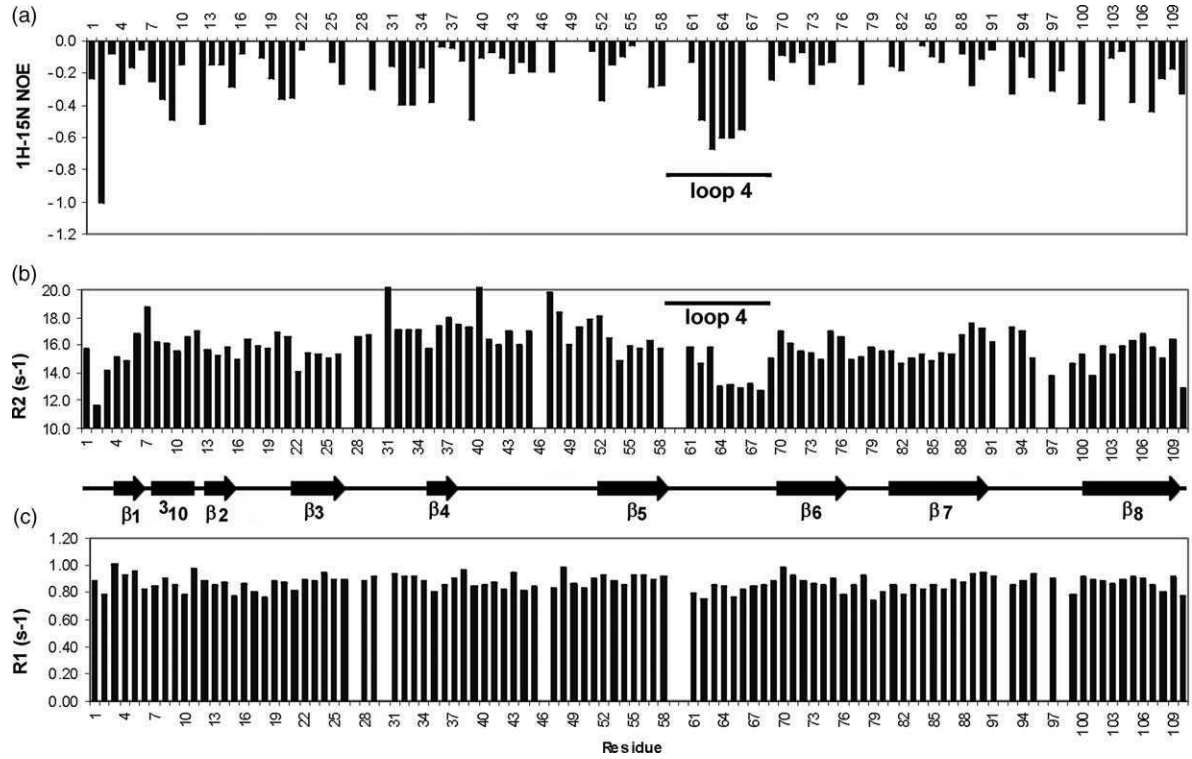

Figure 5. Dynamics from NMR data. (a) ^{1}H–^{15}N heteronuclear NOE data. (b) ^{15}N R_2 relaxation data. (c) ^{15}N R_1 relaxation data. Slowly relaxing, flexible loop L4 is marked. β-Strands, 3_{10}-helix and loops are indicated schematically.

44 μM or no cruzipain (i.e. 60 kDa cruzipain purified from *T. cruzi*). The 2D ^{15}N-HSQC spectra were used to monitor interactions[27] through analysis of chemical shift perturbations (CSPs) and alterations in intensities and broadness of the chagasin signals due to the presence of cruzipain.

Because of the high molecular mass of cruzipain and the high stability of the complex formed ($K_i = 0.095(\pm 0.0076)$ nM at 37 °C; pH 6.5)[17] (slow exchange on the NMR time-scale), only very small CSPs were observed in the presence of cruzipain and the major effect observed upon cruzipain binding to chagasin was the decrease in signal intensity for all chagasin amide signals in a cruzipain concentration-dependent manner. At 22 μM cruzipain, the signal intensity of amide protons decreased by 25% (data not shown), while at 44 μM cruzipain, 50% intensity reduction was observed on average for all chagasin amide signals (Figure 6(a)).

It was, however, curious to note that the intensity reduction was stronger (over a threshold value of 55%; Figure 6(a)) for some residues located at the flexible loop L4 (residues S62, K63, L64, L65, A67 and G69), residues located at loop L2 (residues N29, T31, T32) and the residues R91 and W93 located at loop L6. All these residues map to the same region in the surface of chagasin (Figure 6(b)) suggesting the involvement of loops L2, L4 and L6 in the binding to cruzipain.

Discussion

Structure overview

We used NMR methods[21,22] to determine a high-resolution structure of chagasin in solution and studied its dynamics. The solution structure described here (Figure 3) is in agreement with the previously reported model predicting that chagasin is folded as a single Ig-like domain.[20] However, the alignment and packing of the β-strands and the structure of the loops found in our NMR structure

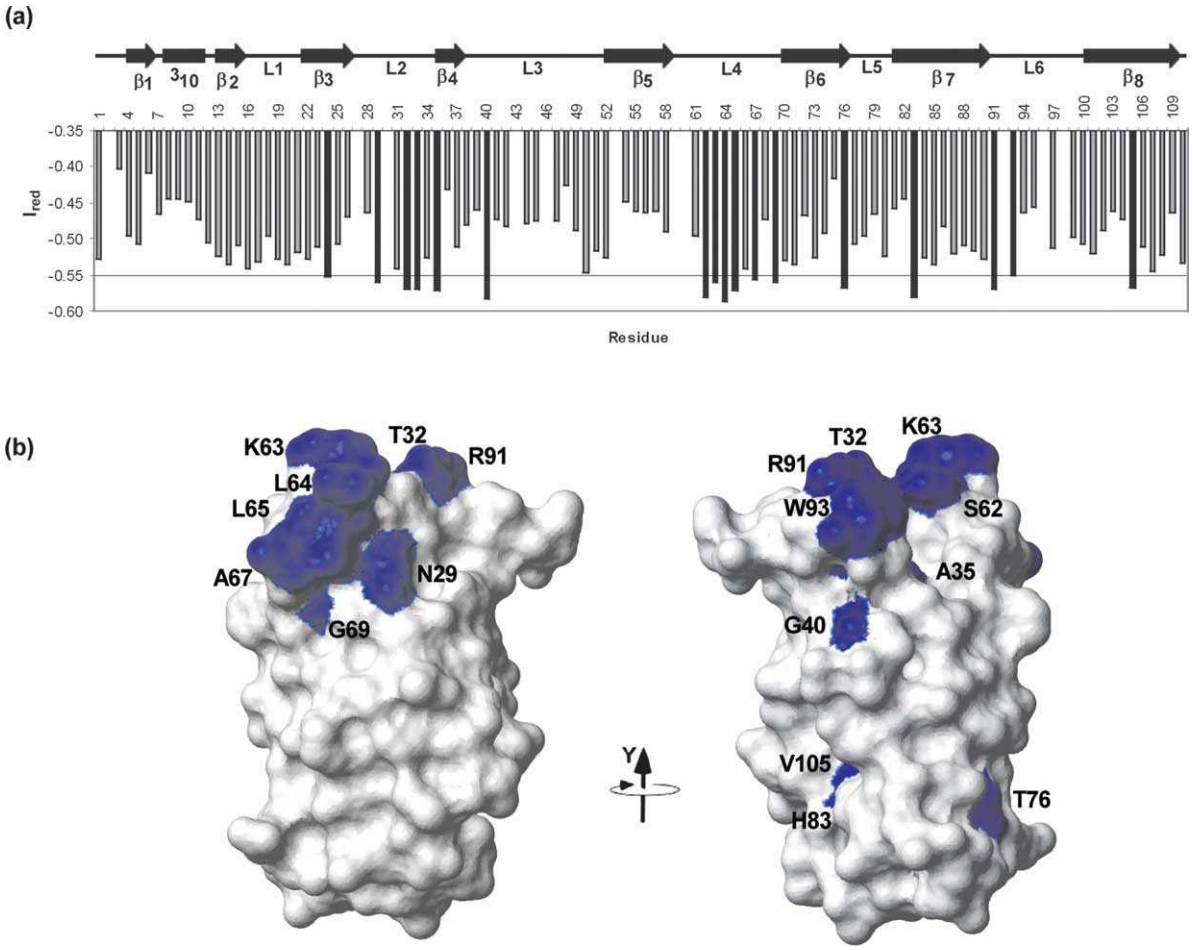

Figure 6. Mapping chagasin binding epitope. (a) ^{1}H–^{15}N HQSC signal intensity reduction upon addition of cruzipain (44 μM) to a sample of ^{15}N-labelled chagasin (100 μM). Chagasin residues showing stronger signal intensity reduction over a threshold of 55% are in black and residues showing smaller signal intensity reduction are shaded gray. β-Strands, 3_{10}-helix and loops are indicated schematically. (b) Residues showing the strongest signal intensity reduction upon cruzipain binding are coloured blue in the molecular surface of chagasin. The surface is in the same orientation as in Figure 3 (left) or rotated by 180° around the *y*-axis (right).

differs significantly from previous models, a finding that is not surprising, considering the remote homology that exists between chagasin and the templates used in the modelling studies.[20] It is noteworthy that the primary sequence of chagasin cloned from the Dm28c *T. cruzi* strain is very similar to those present in the CL Brenner strain.[5] Since most substitutions found between Dm28c and CL Brenner chagasin sequences are located in region of loops (residues in blue in Figure 1), the folding of the corresponding proteins are likely the same.

Regarding the different topological subtypes of the Ig fold,[28,29] the chagasin structure does not seem to match exactly any of the four subtypes previously characterized[29] and thus should be classified as an Ig-like domain. Regarding the number of strands chagasin is somewhat simpler than the Ig type v domain but more complex (higher number of strands) than a typical Ig type c domain.[28,29]

Similarity to other Ig folds

Using our NMR structure as input and the program DALI,[30] which calculates a geometrical similarity score (*Z*-score), we searched for structurally similar Ig-like domains in the Protein Data Bank. Proteins sharing structural similarity to chagasin (*Z*-scores from 5.8 to 2.5) included fragments of human and murine immunoglobulins, cytokine receptors, e.g. IL-1β, IL-4, IL-6; cell adhesion proteins, e.g. Vcam-1, axonin-1, epithelial cadherin, T-cell surface glycoproteins, e.g. CD1 and CD8 and hydrolases e.g. protein-arginine deiminase type IV and sialidase. These proteins all belong to the Ig superfamily, and are involved in protein–protein interactions.

These structures align with the chagasin NMR structure with backbone rmsd values of the overlapped regions in the range of 2.5 Å to 4.0 Å. Figure 1 shows the sequence alignment of several chagasin-like proteins and three human proteins obtained from the DALI searches; namely, the N-terminal domains of the human T-cell surface glycoprotein CD8 α-chain, interleukin-1 type 1 receptor and the vascular cell adhesion molecule 1. Comparison of these sequences shows a number of conserved aromatic or hydrophobic residues. These hydrophobic residues occur often in the β-strand regions and in an alternating manner, so that hydrophobic residues are buried forming a hydrophobic cluster and hydrophilic residues are exposed to the solvent in the three-dimensional structure. Some of the hydrophobic side-chains buried in the chagasin structure are shown in Figure 3, including the residues Leu26 and Leu87, which give rise to very high-field chemical shifts (−1.28, −0.65 for Leu26 and −0.81, −0.46 for Leu87 methyl groups) due to the anisotropic effect of near aromatic rings (e.g. Trp36). This pattern of hydrophobic residue conservation is characteristic of the immunoglobulin domain fold and is important for its stability and thus is present in all structures resulting from the DALI searches. Chagasin-like ICPs do not contain the conserved cysteine residues involved in disulphide bonds, which contribute to the structural stability of Ig. Instead two aliphatic residues Leu27 and Leu87 contribute to the hydrophobic core together with other conserved aromatic residues near the termini of the β-strands (F/Y34, F39, W36, Y/F57, Y89, F/Y103).

Potential binding interface

Some residues in loops L2 (Ser28 to Gly33), L4 (Ler64 to Gly69) and L6 (Arg91 to Trp93) are conserved among chagasin-like proteins of several pathogens (residues displayed in bold; Figure 1). L2 and L6 are equivalent to the so-called complementarity determining regions (CDR 1 and 3, respectively) of the immunoglobulin variable chain domain (e.g. CD8-α; Figure 1) and are supposed to be responsible for chagasin function.[14,15] Moreover, it was recently shown[14] that a peptide with sequence GNPTTGF present in loop L2 (based on the chagasin-like protein of *Entamoeba*), is able to inhibit papain (although, with much higher K_i value than the intact ICP), suggesting the potential importance of this loop for the inhibitory activity.

From our structure, we conclude that the N-terminal β-strand (β1) is not involved in the interaction with cruzipain. In the solution structure β1 is contacting β3 and thus the N terminus is near the C terminus (Figure 3). This result contrasts with the earlier model,[15,20] which described β1 antiparallel to β8 and from which it was predicted that the N terminus was on the opposite side of the structure as the C terminus, near loops L2, L4 and L6, potentially interacting with cruzipain.

Here we attempted to map the interaction site of chagasin with cruzipain monitoring changes in the ^{15}N HSQC spectrum of chagasin upon addition of cruzipain. NMR techniques have been proven to be useful for mapping protein–protein interactions. In most favourable conditions, the molecule being observed by NMR (e.g. a ^{15}N-labelled protein in the complex), are in fast exchange between both the bound and free states. This ideal condition is met by weak interactions with dissociation constants in the range of millimolar to sub-micromolar to sub-μM. In this condition the spectra measured show only one set of signals with populational averaged properties (e.g. chemical shifts and relaxation rates). The relative populations are determined by the dissociation constant and the concentrations of the proteins in the NMR sample. If the free state is more populated then the spectra resemble the free state, but some signals perturbations can be detected, mainly for residues at the binding epitope. This is because residues at the binding epitope usually display markedly different properties in the bound and free states due to environment changes upon binding.

The complex formed between cruzipain–chagasin was determined to be very tight, as measured by the inhibition of the cruzipain proteolytic activity

by chagasin ($K_i = 0.095(\pm 0.0076)$ nM at 37 °C; pH 6.5).[17] Moreover, the native cruzipain purified from *T. cruzi* and used in our studies is a large molecule with molecular mass around 60 kDa, chagasin itself is a 12 kDa protein. Considering these constants and the chagasin/cruzipain molar ratios used in our experiments (100:44 μM) one should expect to have 44 μM chagasin in the bound form (i.e. as the 72 kDa chagasin–cruzipain complex) and 66 μM chagasin in the free form. Slow the exchange between the two forms should occur. As a result, the ^{1}H–^{15}N HSQC spectrum of the mixture showed a set of signals very similar to the free chagasin, but with losses in signal intensities as compared to 100 μM chagasin alone (Figure 6(a)). A second set of signals due to the bound form of chagasin were too broad to be observed.

Theoretically, all signals observed in the ^{1}H–^{15}N HSQC spectrum of the mixture should have their intensities decreased by the same amount (i.e. 44%) if there is any exchange measurable (in the NMR time-scale) between the free and bound states. In practice, an intensity decrease has shown a residue-specific pattern. Although most signals show a basal intensity reduction of 50% on average, for some residues located at the flexible loop L4 (residues S62, K63, L64, L65, A67 and G69), residues located at loop L2 (residues N29, T31, T32) and the residues R91 and W93 located at loop L6, the reduction of signal intensity was more pronounced (over a threshold of 55%; Figure 6(a)).

The residue-specific signal intensity reduction observed in Figure 6(a) could be related to a small contribution of the bound state and thus further evidence of the participation of loops L2, L4 and L6 in the binding to cruzipain.

On the other hand, noise caused by the basal intensity reduction observed for all residues, as well as possible differential modification of relaxation properties of loop and core amide groups unrelated to the binding epitope by anisotropic tumbling upon formation of the large 72 kDa complex, complicate the interpretation of the data and these results should not be overestimated. Thus, more direct experiments (e.g. site-directed mutagenesis) are necessary to more reliably test the involvement of residues located at loops L2, L4 and L6 in the binding to cruzipain.

Conservation with other inhibitors

Our epitope mapping results as well as studies from others[14,15] suggest that chagasin interacts with its ligand by the combination of three loops that form the ligand-binding site. This situation is similar to other CP inhibitors, involving reversible tight-binding interactions. In the cystatin superfamily, the inhibitor is able to block the active site of the protease through two hairpin loops in a such way that neither of its peptidic bonds is in direct contact with the protease catalytic site (steric blockage).[31] It remains to be clarified if chagasin interacts with target CPs in a manner similar to that of cystatin, or if alternatively, a particular loop interacts directly with the enzyme's active site.

Interestingly, it appears that in the conserved CDR2 region of CD8 the sequence SNPTSG is found conserved in all chagasin-like proteins (Figure 1). Mutational studies of CD8 have demonstrated the involvement of both CDR1 and CDR2 loops in MHC class I recognition.[32] In particular, three mutants with non-conservative substitutions located in the LLSNP peptide of the CDR1 loop had the greatest effect on binding to MHC class I. Because human CD8 molecules function as co-receptor on cytotoxic T lymphocytes (CTL), interacting with a non-polymorphic region of the HLA class I α3 domain on antigen-presenting cells, we speculate that chagasin might interfere with the immunomodulation of the CTL immune response by sharing the surface conformational epitope with CD8, modulating the parasite virulence. However, further studies are necessary to verify whether chagasin is really able to interfere with CD8 function.

Backbone dynamics

Finally, $T1$, $T2$, ^{1}H–^{15}N NOE (Figure 5) and $H/^{2}H$ exchange measurements gave insight in the dynamics of the chagasin structure. Several amide protons are involved in strong and stable hydrogen bonding and were found to exchange slowly, in several hours or days (Figure 2). However, most amide protons from β1 and β5 strands were rapidly exchanged by deuterium, hence hydrogen bonds from these strands should contribute less to the overall stability of the fold. The interactions between β1 and β3, and between β5 and β6 are possibly native contacts formed late during folding and one of the first native contacts lost during unfolding.

By analysis of the relaxation rates and ^{1}H–^{15}N NOE for global and internal motions, the global correlation time for isotropic tumbling was estimated to be 13 ns. This is a quite high value for a protein of 12 kDa in a monomeric state in solution and indicates that chagasin is in rapid equilibrium between the monomeric and dimeric forms at a concentration of 0.5 mM. The association constant is estimated to be in the millimolar range as a much smaller R_2/R_1 ratio, compatible with a purely monomeric state, can be measured at a concentration of 0.1 mM (data not shown). Order parameters (S^2) with values between 0.8 and 0.9 were obtained for most amino acids; these values are typical of structured portions in folded globular proteins. In contrast, the different relaxation behavior of residues in loop 4 and the two serine residues located at the N terminus could be fitted by S^2 values between 0.5 and 0.75 and effective correlation times in the sub-nanosecond time-scale.

We concluded that the loop 4 is genuinely flexible in solution. This is also one of the loops implicated in the binding to cruzipain. Flexibility in this loop

could contribute to a better fit of chagasin to the cysteine protease.

In conclusion, the data reported here detail features of the chagasin structure and dynamics and provide experimental information for further understanding the inhibition of CPs by chagasin-like inhibitors.

Materials and Methods

Expression and purification of chagasin

A DNA fragment spanning the entire chagasin open reading frame (ORF) but lacking the ATG initiation codon was amplified by PCR (Expand Long Template PCR System; Roche) with genomic DNA from Dm28c *T. cruzi* as template using the forward oligonucleotide: 5′-CGGGATCCTCCCACAAGGTGACGAAAGC-3′ (where BamHI site is underlined) and the reverse oligonucleotide 5′-CGCTGCAGTCAGTTTGCCTTGAGATATAC-3′ (PstI site is underlined). The PCR fragment of 346 bp was sub-cloned into pT7-blue 3 vector (Novagen) to be fully sequenced. The BamHI/EcoRI fragment of recombinant pT7-chagasin was sub-cloned into BamHI/EcoRI digested pGEX-4T-1 vector (Amersham Biosciences). The fusion construct was sequenced to ensure that the sequence derived from the chagasin gene was in frame with the thrombin-cleavable glutathione *S*-transferase (GST) tag. Chagasin was co-expressed as a GST fusion protein in *Escherichia coli* BL21 (DE3) pLys S. Cells were grown at 37 °C in LB medium containing 60 μg/ml of carbenicilin and 34 μg/ml of chloranphenicol. Expression was induced by addition of 1 mM IPTG and cells were grown for an additional 15 h at 22 °C before centrifugation for 10 min at 6000 rpm at 4 °C, washing with buffer A (50 mM Tris–HCl (pH 8.0), 150 mM NaCl) containing Benzonase (Novagen) and the protease inhibitor cocktail Complete (Roche), and lysed with a French Press (2×1000 bar). GST-chagasin was then purified by affinity chromatography on a glutathione Sepharose column and eluted with 10 mM reduced glutathione in buffer A. This eluate was then dialyzed against buffer A while the GST moiety was cleaved with thrombin (Roche) (100 units/1 mg of oGST-chagasin) and then purified again on a glutathione Sepharose column. The free chagasin found in the flow-through was concentrated and separated by size-exclusion chromatography on a Superdex 75 column (Amersham Biosciences) to yield chagasin with an additional N-terminal glycine and serine residue coming from the thrombin-cleavable GST tag. All chromatography steps were performed at 4 °C. Uniformly ^{15}N and ^{13}C, ^{15}N-labelled chagasin were grown in *E. coli* BL21(DE3) pLys-S cells in M9 minimal medium containing 0.5 g/l of $^{15}NH_4Cl$ and either 1.3% (w/v) [$^{12}C_6$]glucose or 0.5% (w/v) [$^{13}C_6$]glucose, respectively, as the sole nitrogen and carbon sources, and purified as described above. Protein molecular masses were confirmed by mass spectrometry.

NMR spectroscopy

All NMR experiments[21,22] for ^{1}H, ^{13}C and ^{15}N chemical-shift assignments and structure determination were acquired at 298 K on chagasin dissolved in 20 mM phosphate buffer (pH 5.6) 50 mM NaCl and 0.02% (w/v) sodium azide. The 2D-NOESY (100 ms mixing time) and total correlation spectroscopy (TOCSY) (35 and 70 ms spin-lock) spectra were acquired in both 90% H_2O/10% $^{2}H_2O$ and in 99.98% $^{2}H_2O$ using a 1.19 mM unlabelled chagasin sample. A 0.52 mM ^{15}N-labelled sample in 90% H_2O (10% $^{2}H_2O$) was used for ^{15}N-HSQC and 3D ^{15}N-edited NOESY experiments. 0.73 mM ^{13}C,^{15}N-labelled sample was used for the 3D HCCH COSY and ^{13}C-edited NOESY experiments. Another ^{13}C,^{15}N-labelled sample 0.3 mM was used to acquire 3D CBCA(CO)NNH, CBCANNH, HBHA(CO)NNH, H(CCCO)NH-TOCSY (13 ms), (H)CC(CO)NH-TOCSY (13 ms) and HNCO experiments. Spectra were acquired in a Bruker DRX600 spectrometer equipped with a cryo-probehead (triple-resonance experiments) or conventional TXI inverse probehead (homonuclear experiments). Data were processed using the XWIN-NMR (v.3.0) (Bruker BioSpin GmbH, Germany). Assignment was carried out using the interactive program SPARKY (v.3.106) (T. D. Goddard & D. G. Kneller, University of California, San Francisco).

NOE assignment and structure calculation

Assignment of NOESY spectra and structure calculation was made iteratively using the program ARIA 1.2[23,24] with CNS 1.1.[33] Initially we manually assigned NOEs characteristic of secondary structure involving amide–amide, amide–H^{α} or H^{α}–H^{α} protons and some unambiguous side-chain interactions. Hydrogen bonds were assigned from slow-exchanging amide protons identified in a 2D ^{15}N HSQC spectrum following exchange into $^{2}H_2O$. Further NOEs were assigned automatically by ARIA. For the structure calculations we used also phi and psi-dihedral angles derived form CSI.[34] In the last ARIA iteration 200 structures were calculated by restrained simulated annealing and the 15 best structures regarding total energy were refined in an explicit water-box and taken as representative ensemble.

Relaxation rates and ^{1}H–^{15}N NOEs

^{15}N T_1 and T_2 relaxation times were extracted from two series of eight 2D ^{1}H–^{15}N correlated spectra with relaxation delays of 12, 52, 102, 202, 402, 902, 2002, and 4002 ms for T_1 and 6, 10, 18, 34, 82, 162, 202, and 242 ms for T_2. Rates were fitted as implemented in SPARKY (v.3.106) (T. D. Goddard & D. G. Kneller, University of California, San Francisco) and plotted as relaxation rates $R_1 = 1/T_1$ and $R_2 = 1/T_2$. Steady-state ^{1}H–^{15}N NOEs were determined according to the formula $\mathrm{NOE} = (I - I_{\mathrm{ref}})/I_{\mathrm{ref}}$, where I is intensity of a cross-peak in a 2D ^{1}H–^{15}N correlated spectrum with broadband ^{1}H presaturation and I_{ref} is the intensity in a reference spectrum recorded without pre-saturation. All experiments were performed on a 0.52 mM ^{15}N-labelled chagasin sample.

Mapping of the chagasin:cruzipain interaction

Native cruzipain (60 kDa) was purified from *T. cruzi* Dm28c epimastigotes as described.[35] After purification to homogeneity, the protein was further concentrated by centrifugation at 3000*g* using Centricon (Amicon, Bedford, MA) filters with a cut-off of 50 kDa. Three samples were prepared containing, respectively, 100 μM ^{15}N-labelled chagasin alone, or in the presence of 22 or 44 μM cruzipain in 20 mM phosphate buffer (pH 5.6),

50 mM NaCl and 0.02% sodium azide. The 2D ^{15}N-HSQC spectra were acquired with 4096×1024 points and 16 scans. Relative signal intensity reduction was calculated according to the formula $I_{red} = (I - I_{ref})/I_{ref}$, where I is the intensity of a chagasin cross-peak in the presence of cruzipain and I_{ref} is the intensity of the cross-peak in the reference sample of chagasin alone. I_{red} was plotted against the residue number.

Atomic coordinates, NMR restraints and chemical shift assignments

The atomic coordinates and NMR restraints for the ensemble of the 15 best structures calculated for chagasin were added to the RCSB Protein Data Bank, PDB entry 2FO8. Chagasin chemical shift assignments were deposited in the BMRB, entry 6876.

Acknowledgements

This work was supported financially by FAPERJ and CNPq (Brazilian Funding agencies). J.S. and A.P.C.A.L. acknowledge the Wellcome Trust for funding. We thank E. Pays (Brussels) for critical reading of the manuscript.

References

1. Meirelles, M. N., Juliano, L., Carmona, E., Silva, S. G., Costa, E. M., Murta, A. C. & Scharfstein, J. (1992). Inhibitors of the major cysteinyl proteinase (GP57/51) impair host cell invasion and arrest the intracellular development of *Trypanosoma cruzi in vitro*. *Mol. Biochem. Parasitol.* **52**, 175–184.
2. McGrath, M. E., Eakin, A. E., Engel, J. C., McKerrow, J. H., Craik, C. S. & Fletterick, R. J. (1995). The crystal structure of cruzain: a therapeutic target for Chagas' disease. *J. Mol. Biol.* **247**, 251–259.
3. Engel, J. C., Doyle, P. S., Hsieh, I. & McKerrow, J. H. (1998). Cysteine protease inhibitors cure an experimental *Trypanosoma cruzi* infection. *J. Expt. Med.* **188**, 725–734.
4. Cazzulo, J. J. (2002). Proteinases of *Trypanosoma cruzi*: potential targets for the chemotherapy of Chagas disease. *Curr. Top. Med. Chem.* **2**, 1261–1271.
5. El-Sayed, N. M., Myler, P. J., Bartholomeu, D. C., Nilsson, D., Aggarwal, G., Tran, A. N. *et al*. (2005). The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science*, **309**, 409–415.
6. Eakin, A. E., Mills, A. A., Harth, G., McKerrow, J. H. & Craik, C. S. (1992). The sequence, organization and expression of the major cysteine proteinase (cruzain) from *Trypanosoma cruzi*. *J. Biol. Chem.* **267**, 7411–7420.
7. Huete-Perez, J. A., Engel, J. C., Brinen, L. S., Mottram, J. C. & McKerrow, J. H. (1999). Protease trafficking in two primitive eukaryotes is mediated by a prodomain protein motif. *J. Biol. Chem.* **274**, 16249–16256.
8. Bonaldo, M. C., d'Escoffier, L. N., Salles, J. M. & Goldenberg, S. (1991). Characterization and expression of proteases during *Trypanosoma cruzi* metacyclogenesis. *Expt. Parasitol.* **73**, 44–51.
9. Sajid, M. & McKerrow, J. H. (2002). Cysteine proteases of parasitic organisms. *Mol. Biochem. Parasitol.* **120**, 1–21.
10. Scharfstein, J., Schmitz, V., Morandi, V., Capella, M. M., Lima, A. P., Morrot, A. *et al*. (2000). Host cell invasion by *Trypanosoma cruzi* is potentiated by activation of bradykinin B(2) receptors. *J. Expt. Med.* **192**, 1289–1300.
11. Aparicio, I. M., Scharfstein, J. & Lima, A. P. (2004). A new cruzipain-mediated pathway of human cell invasion by *Trypanosoma cruzi* requires trypomastigote membranes. *Infect. Immun.* **72**, 5892–5902.
12. Rawlings, N. D., Tolle, D. P. & Barrett, A. J. (2004). Evolutionary families of peptidase inhibitors. *Biochem. J.* **378**, 705–716.
13. Ivens, A. C., Peacock, C. S., Worthey, E. A., Murphy, L., Aggarwal, G., Berriman, M. *et al*. (2005). The genome of the kinetoplastid parasite, *Leishmania major*. *Science*, **309**, 436–442.
14. Riekenberg, S., Witjes, B., Šarić, M., Bruchhaus, I. & Scholze, H. (2005). Identification of EhICP1, a chagasin-like cysteine protease inhibitor of *Entamoeba histolytica*. *FEBS Letters*, **579**, 1573–1578.
15. Rigden, D. L., Mosolov, V. V. & Galperin, M. Y. (2002). Sequence conservation in the chagasin family suggests a common trend in cysteine proteinase binding by unrelated protein inhibitors. *Protein Sci.* **11**, 1971–1977.
16. Sanderson, S. J., Westrop, G. D., Scharfstein, J., Mottram, J. C. & Coombs, G. H. (2003). Functional conservation of a natural cysteine peptidase inhibitor in protozoan and bacterial pathogens. *FEBS Letters*, **542**, 12–16.
17. Monteiro, A. C., Abrahamson, M., Lima, A. P., Vannier-Santos, M. A. & Scharfstein, J. (2001). Identification, characterization and localization of chagasin, a tight-binding cysteine protease inhibitor in *Trypanosoma cruzi*. *J. Cell Sci.* **114**, 3933–3942.
18. Santos, C. C., Sant'anna, C., Terres, A., Cunha-e-Silva, N. L., Scharfstein, J. & Lima, A. P. (2005). Chagasin, the endogenous cysteine-protease inhibitor of *Trypanosoma cruzi*, modulates parasite differentiation and invasion of mammalian cells. *J. Cell Sci.* **118**, 901–915.
19. Besteiro, S., Coombs, G. H. & Mottram, J. C. (2004). A potential role for ICP, a Leishmanial inhibitor of cysteine peptidases, in the interaction between host and parasite. *Mol. Microbiol.* **54**, 1224–1236.
20. Rigden, D. L., Monteiro, A. C. S. & de Sá, M. F. G. (2001). The protease inhibitor chagasin of *Trypanosoma cruzi* adopts an immunoglobulin-type fold and may have arisen by horizontal gene transfer. *FEBS Letters*, **504**, 41–44.
21. Sattler, M., Schleucher, J. & Griesinger, C. (1999). Heteronuclear multidimensional NMR experiments for the determination of proteins in solution employing pulsed field gradients. *Prog. NMR Spectrosc.* **34**, 93–158.
22. Kay, L. E. (1995). Pulsed field gradient multidimensional NMR methods for the study of protein structure and dynamics in solution. *Prog. Biophys. Mol. Biol.* **63**, 277–299.
23. Nilges, M. & O'Donoghue, S. I. (1998). Ambiguous NOEs and automated NOE assignment. *Prog. NMR Spectrosc.* **32**, 107–139.
24. Linge, J. P., O'Donoghue, S. I. & Nilges, M. (2001). Automated assignment of ambiguous nuclear Overhauser effects with ARIA. *Methods Enzymol.* **339**, 71–90.
25. Clore, G. M., Szabo, A., Bax, A., Kay, L. E., Driscoll, P. C. & Gronenborn, A. M. (1990). Deviations from the

simple 2 parameter model free approach to the interpretation of ^{15}N nuclear magnetic relaxation of protein. *J. Am. Chem. Soc.* **112**, 4989–4991.
26. Dosset, P., Hus, J.-C., Blackledge, M. & Marion, D. J. (2000). Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data. *J. Biomol. NMR*, **16**, 23–28.
27. Hajduk, P. J., Meadows, R. P. & Fesik, S. W. (1999). NMR-based secreening in drug discovery. *Quart. Rev. Biophys.* **32**, 211–240.
28. Bork, P., Holm, L. & Sander, C. (1994). The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.* **242**, 309–320.
29. Potapov, V., Sobolev, V., Edelman, M., Kister, A. & Gelfand, I. (2004). Protein–protein recognition: juxtaposition of domain and interface cores in immunoglobulins and other sandwich-like proteins. *J. Mol. Biol.* **342**, 665–679.
30. Holm, L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**, 316–319.
31. Otlewski, J., Jelen, F., Zakrzewska, M. & Olesky, A. (2005). The many faces of protease-protein inhibitor interaction. *EMBO J.* **24**, 1303–1310.
32. Sanders, S. K., Fox, R. O. & Kavathas, P. (1991). Mutations in CD8 affect interactions with HLA class I and monoclonal anti-CD8 antibodies. *J. Expt. Med.* **174**, 371–379.
33. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W. *et al*. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallog. sect. D*, **54**, 905–921.
34. Wishart, D. S. & Sykes, B. D. (1994). The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J. Biomol. NMR*, **4**, 171–180.
35. Murta, A. C. M., Persechini, P. M., de Souto Padrón, T., de Souza, W., Guimarães, J. A. & Scharfstein, J. (1990). Structural and functional identification of GP57/51 antigen of *Trypanosoma cruzi* as a cysteine proteinase. *Mol. Biochem. Parasitol.* **43**, 27–38.

Edited by M. F. Summers

(Received 27 October 2005; received in revised form 13 January 2006; accepted 17 January 2006)
Available online 3 February 2006