

# Caracterización de la *Familia de Ecm33* como etapa en la predicción de la estructura y función de Ecm33

David Jiménez-Morales

Departamento de Microbiología II. Facultad de Farmacia. Universidad Complutense de Madrid (UCM)

## Introducción

Ecm33p/YBR078W (SaEcm33p) es una proteína de 469 aminoácidos de anclaje GPI perteneciente a la **Familia de Ecm33** [1] (anteriormente llamada, *Familia de SPS2* [2]) formada en *Saccharomyces cerevisiae* por un conjunto de proteínas homólogas que son, PSTI/YDR055W, SPS2/YDL052 y SPS22/YCL048W. Todas ellas comparten características típicas de proteínas de anclaje a GPI, es decir, tienen péptido señal, una región de la secuencia rica en serina y treonina, junto con un dominio potencial C-terminal de anclaje a GPI [3]. A este respecto, en el caso de SaEcm33p, se ha demostrado la presencia en su secuencia de la región *w-minus* que determina el anclaje GPI a la membrana plasmática, de tal forma que se ha conseguido determinar el anclaje a membrana o pared modificando esta región [4].

Aunque la función molecular exacta de este grupo de proteínas es desconocida, diferentes evidencias experimentales la relacionan en general con la integridad de la pared celular.

- En el caso de Ecm33p, el mutante de levadura *ecm33Δ* es termosensible, incrementa la sensibilidad a calcofluor y estrés oxidativo [5], [6], [7]. Además presenta un crecimiento desorganizado de la pared celular [1].
- Sobre *PST1* fue publicado que su expresión se produce en diferentes cepas mutantes para la pared celular o en respuesta a daño en la pared celular transitorio [8], [9], actuando en el mecanismo compensatorio que desencadena la cascada de Slt2p-MAP quinasa responsable de la integridad de la pared celular. La cepa *pst1Δ* de levadura no presenta un fenotipo distinguible.
- Por su parte, *SPS2* es secretada en las etapas tardías de la meiosis [10], y parece estar relacionada con la síntesis de la pared de la espora.
- *SPS22* es requerida redundantemente junto con *SPS2* en la organización de la capa de Beta-glucano de la pared de la espora.

Algunas de estas proteínas de la familia de Ecm33 han sido identificadas en *Candida albicans*, como es el caso de CaEcm33p (Ecm33.3f|CA3115 [1]). La delección del gen *ECM33* en *Candida* tiene como consecuencia defectos en la pared celular y superficie. Además es requerida para la filamentación normal *in vitro* y se ha visto que presenta un efecto dosis dependiente. También fue relacionada su implicación en la virulencia [1]. Otras dos proteínas de la familia de Ecm33 en *Candida albicans* son CA2181|ECM331 (homóloga de Pst1) y CA0513|IPF13972 (alta homología con SaSps22), aunque no existen datos experimentales sobre ellas. El hecho de que no exista el homólogo de SaSps2p está en concordancia, en principio, con la ausencia de meiosis en el hongo patógeno.

Puesto que la función molecular de este conjunto de proteínas es desconocida, nos hemos propuesto el empleo de diferentes herramientas y estrategias bioinformáticas que, con carácter predictivo, puedan orientar la posterior investigación científica. Para abordar esta tarea nos decantamos por

establecer como proteína-referencia de la familia a la propia Ecm33. Entre los motivos se encuentra el hecho de que la delección de este gen, aunque viable, presenta una manifestación fenotípica de interés: la desorganización de la pared celular de la levadura (Imagen 1). Además, Ecm33 es capaz de complementar las mutantes de *fsr2-1* y *las21Δ*, los cuales presentan defectos en la síntesis de GPI [6]. A este respecto y con anterioridad se había demostrado que Las1 no controla la expresión de Ecm33 [11].

Una de las causas en la dificultad para poder detectar una función es consecuencia de que no presenta homologías significativas con proteínas de función conocida. Un primer intento en tratar de atribuirle una función lo hicieron Lussier et al, los cuales sugirieron que podría tener una función en la síntesis de los polisacáridos de la pared celular [5].

En el presente trabajo estudiaremos las diferencias existentes entre los diferentes miembros de la Familia Ecm33, teniendo en cuenta el análisis de las dos proteínas Ecm33, tanto de *S. cerevisiae* como en *C. albicans*, ya que podemos considerar significativas las diferencias existentes entre ambas [469aa y 423aa; Score = 159 bits (401), Expect = 2e-38 Identities = 97/331 (29%), Positives = 160/331 (48%), Gaps = 6/331 (1%)] . A pesar de esto, la función molecular parece ser la misma puesto que la expresión en el mutante de levadura *ecm33Δ* del gen *CaECM33* complementa los defectos en la pared celular que presentaba el mutante *ecm33Δ*. En lo que respecta a Pst1p, a pesar de que el mutante *pst1Δ* carece de fenotipo apreciable en *S. cerevisiae*, la delección de *pst1* en el mutante *ecm33Δ*, incrementa las anomalías en la pared celular.

Por tanto, uno de los objetivos es emplear estrategias bioinformáticas encaminadas a la predicción de la función molecular, como la caracterización a nivel de secuencia de la Familia de Ecm33; el estudio de las secuencias a diferentes niveles (características físico-químicas, composición, exposición a solvente, estructura secundaria...) junto con el empleo de los métodos más adecuados de cara a la predicción de la estructura tridimensional de Ecm33p.

En este punto trataremos de servirnos del paradigma de la predicción de la función mediante la estructura en proteínas de función desconocida –el también llamado paradigma de la “secuencia-a-estructura-de-estructura-a-función”. Esta paradigma está basado en el hecho de asumir que los patrones de estructura tridimensional están conservados a lo largo de una mayor distancia evolutiva que patrones reconocibles en la secuencia primaria [12]. Y esta asunción a su vez está basada en los datos extraídos de la base de datos de estructuras proteicas (PDB) que muestran plegamientos similares para proteínas con grandes distancias proteicas [12], [13], [14], [15], [16].

Todo ello para tratar de explicar la presencia de las proteínas de la Familia de Ecm33 en la superficie celular y su relación a su vez con la pared celular de los organismos fúngicos en los que han sido identificadas, especialmente, *S. cerevisiae* y *C. albicans*.

## Métodos

### Bases de datos

NCBI Non Redundant, nr, pdbaa, yeastaa  
Protein Data Bank  
Swiss-Prot Release 44.7 of 11-Oct-2004  
Candida Data Release R2 (Feb 9, 2004)  
S.pombe database (Trust Sanger Institute <http://www.sanger.ac.uk/>)  
A. fumigatus (Trust Sanger Institute <http://www.sanger.ac.uk/>)

Los datos de interacciones proteicas fueron obtenidos de BIND (<http://bind.ca/>), DIP (<http://dip.doe-mbi.ucla.edu/dip/>), GRID ([http://biodata.mshri.on.ca/yeast\\_grid/](http://biodata.mshri.on.ca/yeast_grid/)) o PathCalling ([http://portal.curagen.com/pathcalling\\_portal/index.htm](http://portal.curagen.com/pathcalling_portal/index.htm)).

Se emplearon diferentes algoritmos para la búsqueda de homólogos de la familia de Ecm33, BLAST, PSI-BLAST [17] y FASTA [18]. Los alineamientos múltiples entre las proteínas seleccionadas fueron realizados con T-Coffee [19] y con HMMer [20] se construyeron los Modelos Ocultos de Markov (HMM).

En la construcción el árbol filogenético se empleó *Evolutionary Trace Server* (TraceSuite II: <http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html>) [21]

En los análisis preliminares se llevaron a cabo búsquedas en las diferentes base de datos de familias, dominios, sitios funcionales y demás herramientas de análisis de secuencias tales como

- PFAM (<http://www.sanger.ac.uk/Software/Pfam>),
- CDD (<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>)
- InterPro (<http://www.ebi.ac.uk/interpro>),
- BLOCKS (<http://blocks.fhcrc.org/blocks/>),
- PRINTS (<http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>),
- SMART (<http://smart.embl-heidelberg.de/>)
- ProDom (<http://protein.toulouse.inra.fr/prodom/current/html/home.php>),
- Prosite (<http://us.expasy.org/prosite/>).
- MEME & MAST (<http://meme.sdsc.edu/meme/website/>)

Para la predicción de la presencia de péptido señal y localización celular de las diferentes proteínas se emplearon los servidores de PSORT II (<http://psort.ims.u-tokyo.ac.jp/>) y Signal IP [22]. En la predicción de proteínas de anclaje a GPI se empleó big-PI Predictor [23].

En el estudio y análisis de las composición y características de la secuencia, utilizamos ProtParam Tool (<http://www.expasy.org/tools/protparam.html>).

En la predicción de la estructura secundaria tuvimos en cuenta los datos obtenidos a través de PSIPRED [24] y PHD-PredictProtein ([25], además de Jufo, Jufo3D y SAM, todos ellos disponibles en el servidor automático de predicción de estructuras Robetta. El conjunto de resultados de las predicciones de estructuras secundarias, junto con los datos de los alineamientos múltiples, orientó el proceso que seguimos para dividir la secuencia de SaEcm33.

Para la predicción de la estructura tridimensional y basándonos en los resultados de PSI-BLAST frente a la base de datos de estructuras, consideramos más adecuado el empleo de servidores de Fold Recognition, o aquellos que modelen *de novo*. Los análisis de Fold Recognition fueron

realizados con FFAS [26] [27], mientras que los modelos fueron obtenidos mediante 3D-PSSM [28] y LOOPP [29]. Para el empleo de métodos de modelado *de novo* utilizamos el servidor automático Robetta [30], que combina en función de la secuencia remitida, el modelado comparativo y el modelado *de novo*.

Robetta sigue varios pasos en su proceso predictivo, comenzando por el empleo de un procedimiento de investigación jerarquizado, llamado *Ginzu*, con el objetivo de identificar los posibles dominios de la secuencia. Le sigue el empleo de *K\*Sync*, un nuevo método para el alineamiento de una secuencia problema frente a la secuencia-estructura de un parental seleccionado, seguido a continuación por el *modelado comparativo* [31], que se basa en el método de la inserción de fragmentos *Rosetta*. En el caso de que no existe un modelo parental, Robetta procede al modelado mediante el empleo de una ligera modificación del protocolo para la predicción de estructuras *de novo* [32]. Una vez obtenido los modelos se hacen búsquedas mediante MAMMOTH [33] para la detección de similitudes estructurales entre modelos y estructuras de PDB.

Las imágenes fueron generadas con Rasmol [34] y PyMOL (DeLano Scientific; <http://pymol.sourceforge.net/>)

## Resultados

### Contexto genético

El conocimiento de un determinado gen en otros genomas (Predicción de la Función basado en ortólogos) proporciona más información específica sobre la función proteica, mientras que el análisis de la secuencia en su contexto genómico (Predicción de la Función basado en el contexto) proporciona información sobre su contexto funcional. Con este fin hemos analizado la región de DNA de entre 30-40kb alrededor de los genes motivos de este estudio.

En levadura, lo primero que llama la atención es que *ECM33* y *PST1*, cromosomas II y IV respectivamente, tienen en sus proximidades genes que están relacionados con el **transporte de aminoácidos**, como es *TAT1* en *Ecm33*, mientras que en el caso de *PST1* tiene cerca a *BAP3*, una permeasa relacionada con los mismos aminoácidos. Además, genes implicados en el **transporte intracelular**, como *YOS9* en el cromosoma IV y *SEC18* en el cromosoma II. También se encuentra en el contexto de *ECM33* proteínas necesarias para la **viabilidad celular** (*YBR070C*). Cerca de *PST1* se encuentra *TPH1*, con la misma característica. En ambos entornos podemos encontrar también **genes que median en los procesos de degradación proteica** (*UBC4* en el cromosoma II; *CDC 34* y *UBC5* en el cromosoma IV). En los alrededores de *SPS2*, situado también en el cromosoma IV, principalmente se encuentran genes involucrados en la **meiosis** (*EMI1*, *EMI2*, *SPS1*). También existen **los de transporte intracelular** (*FPR2*) y la **viabilidad celular** (*RBA50*). Algo parecido sucede en el caso de *SPS22*, en cuyas proximidades encontramos a *MRC1* y *KAR4* implicados en la meiosis, o *LRE1*, relacionada con la estructura de la pared celular; *GRX1*, gen de respuesta a stress oxidativo.

Con respecto al estudio de *ECM33* en otros genomas conocidos, como es el de *C. albicans*, encontramos que en sus contextos genéticos se encuentra una mayoría de genes de función aún por determinar. A pesar de esto, en las proximidades de *ECM33.3* (cromosoma I), encontramos, al igual que sucede en la levadura, tres genes implicados en el **transporte intracelular** (*BMH2*, *CDC12*, *GGAI*) y otro implicado en la **degradación** (*COQ4*).

### Interacciones

En trabajos recientes se está comenzando a establecer modelos de predicción de la función basados en los datos disponibles de interacciones [35] [36] [37]. Con respecto a los datos de interacción de las proteínas que estamos estudiando, corresponden a SaEcm33 la mayor colección de datos y es la siguiente:

#### 2-híbridos Levadura

YNL124W (NAF1)

#### BIND

Abf1 - Fkh2 - Mbp1 - Ndd1 - Swi4 - Swi6 - SLT2 - Cse1 - Mlp1 - Mlp2 - Nic96 - Nup100 - Nup116 - Nup2 - Xpo1

#### DIP

HRT1

#### GRID

SLT2 / YHR030C

HRT1 / YOL133W

### Predicciones de Péptido señal y anclaje a Gpi. Motivos

De forma experimental se ha demostrado que *Ecm33* presentan anclaje GPI a membrana, así como la región de su secuencia responsable de la localización de este anclaje en la membrana y no en la pared celular. Es por lo tanto interesante estudiar las predicciones en este sentido para el resto de miembros iniciales de la familia, que son las que se muestran en la **tabla 1**.

En la búsqueda de motivos funcionales de centros activos en proteínas con función catalítica, los resultados fueron negativos. Lo más destacable fue un parecido razonable con el motivo funcional las Proteín Tirosín Fosfatasas. Este motivo tiene la siguiente secuencia consenso: (I/V)HCxxGxxRS(T) [38]. En el caso de SaEcm33 encontramos un motivo entre los aminoácidos 350-359 con la siguiente secuencia: VCknGatSTSV. Esta cisteína, situada en la posición 351 en SaEcm33, se encuentra muy conservada en la familia Ecm33. Si hacemos búsquedas del motivo consenso VCxxGxxST (donde x es cualquier aminoácido), solo dos proteínas presentan dicho motivo: por un lado ECM33 y por otro el Receptor de la Interleukina 6 (IL6A\_HUMAN – ver **imagen 2**). De hecho, la mutación de esta Cisteína (C132A) supone la pérdida total de la unión del ligando [39].

### **Predicción de Asn-glucosilaciones, características físico-químicas, y estructura secundaria**

La modificación post-traducciona estudiada es la glucosilación. Los resultados para los principales componentes de la familia de Ecm33 se muestran en datos complementarios disponibles *on line*.

Los resultados de la predicción de la estructura secundaria apuntan hacia una estructura principalmente beta, y existe un relativo consenso por parte de todos los servidores empleados, y en general, para todas las proteínas analizadas. Un análisis más detallado de la predicción de la estructura secundaria lo llevamos a cabo en SaEcm33 (**Figura 2**), para lo cual se emplearon un mayor número de servidores de predicción de estructura secundaria. Los resultados que obtuvimos nos sugirieron la subdivisión de la secuencia en dos regiones. Esto se veía reforzado a raíz de los datos obtenidos de las búsquedas a través de servidores de familias y dominios proteicos (ver métodos), de tal forma que variaban las predicciones de regiones de SaEcm33p que presentaban homología con el dominio del Receptor-L-Domain, y en función de si la búsqueda se realizaba en PFAM o en CDD.

Tras esta subdivisión, analizamos las características físico-químicas, dedicando especial atención a la composición aminoacídica y a la proporción de aminoácidos con carga positiva (Argininas y Lisinas) y a los cargados negativamente (Aspartato y Glutámico). Cabe destacar que mientras en la región N-terminal existe una mayor cantidad de aminoácidos con carga negativa, en la región C-terminal ocurre lo contrario, es decir, un mayor número de aminoácidos con carga positiva, aunque la diferencia es menor.

Para extrapolar estos análisis al resto de miembros de la familia de Ecm33, dividimos las secuencias en función de los alineamientos múltiples. Como puede observarse en la **tabla 2**, podemos comprobar que los resultados son muy similares para el resto de miembros de la familia y en las proteínas que son seleccionados como parentales en el modelado comparativo y para *threading*, como veremos más adelante.

### **BLAST, PSI-BLAST, HMMER: Caracterización de la Familia Ecm33 a nivel de secuencia.**

Para la caracterización de la familia de Ecm33, y partiendo de los miembros de esta proteína en levadura (esto es, Ecm33p, Pst1, Sps2 y Sps22), comenzamos la identificación de los principales ortólogos en otros organismos. Las búsquedas comenzaron frente a la base de datos de *Candida*, a fin de identificar los homólogos ya de sobra conocidos: Ecm33.3|CaEcm33 (homólogo de SaEcm33), Ecm33.1|CaPst1 (homólogo de Pst1) e IPF13972|CaSps22 (proteína homóloga de Sps22). Todo este conjunto de proteínas, junto con las nuevas que iban siendo identificadas en otros organismos, fueron incorporándose a la construcción de los alineamientos múltiples y los correspondientes HMMs de cara a mejorar las búsquedas.

En la **tabla 3** se muestran los organismos en los cuales han sido identificados los homólogos (cuya mayoría de genomas están aún por secuenciar completamente).

### **Estudio Filogenético**

De cara al estudio filogenético del conjunto de proteínas que forman parte de la Familia de Ecm33 utilizamos el *TraceSuite II Server* (ver Métodos). En el estudio se utilizaron los alineamientos múltiples correspondientes a todos los miembros de la Familia de Ecm33 que hemos identificado (**tabla3**), junto con las regiones C-terminal y N-terminal de los principales miembros de la familia. Los resultados se muestran en la **figura 1**.

### **Predicción de Estructura**

Del análisis de los resultados obtenidos mediante búsquedas frente a la base de datos de estructuras (PDB) a través de métodos sensibles como son HMMer y PSI-Blast, llegamos a la conclusión de que el bajo porcentaje de identidad descartan el empleo de métodos basados en el modelado por homología tal como Swiss-Model [40], CHPmodels [41], 3D-JIGSAW [42] o ESyPred3D [43]. Por tanto, para la predicción y análisis de la estructura proteica nos decidimos por el empleo de servidores de reconocimiento de plegamiento (Fold Recognition, **tabla 4**) y métodos *Ab initio* (ver “**Resultados de Robetta**”).

## Discusión

Con los datos de los genomas en la actualidad disponibles y tras los resultados obtenidos de los métodos de detección de homólogos que hemos empleado, parece evidente que se trata de una familia de proteínas que forma parte exclusivamente de organismos fúngicos. Conforme se vayan ampliando y completando el número de genomas (especialmente de organismos fúngicos), podremos ir detectando un mayor número de proteínas que pertenezcan a la familia de Ecm33.

Tras el análisis de homologías y sobre todo, tras el análisis filogenético realizado en este trabajo, pueden apreciarse dos principales ramas evolutivas, es decir, dos principales tipos proteicos, dos subfamilias dentro de la familia de Ecm33, que son:

- **Proteínas del tipo Ecm33 y Pst1.**
- **Proteínas del tipo Sps2 y Sps22.**

Estas subfamilias se corresponden perfectamente en los datos experimentales que hasta la fecha hay disponibles, puesto que mientras la Subfamilia de Ecm33 y Pst1 se expresa en la célula normal, la subfamilia de Sps22 y Sps2 está relacionada con la formación de la espora.

Aunque no es descartable la posibilidad de que Ecm33 se encuentre formando parte de la pared celular [44], no hay duda del anclaje GPI a la membrana plasmática [4]. Ahora bien, si Terashima et. al. afirmaban que el motivo necesario para la localización en la membrana plasmática de Ecm33 es SKKSK, en vista de que CaEcm33 complementa la función de SaEcm33 [1] es de suponer que, analizando la secuencia de CaEcm33, lo determinante en la región *w-minus* son las dos Lys consecutivas que preceden a la Gly, puesto que solo dos regiones sospechosas encontramos en CaEcm33 (393-SSSKKG ó 364-SSKKSG).

Como ha sido descrito previamente [45], entre las propiedades biológicas que obtiene una proteína que ha sido o-glucosilada, destaca aquella en la que adquiere rigidez y estabilidad proteica. Además se ha visto que la adición de o-glucanos en dominios ricos en Ser/Thr (/Pro) puede ser importantes para un reconocimiento funcional. Estos dominios glucosilados pueden proporcionar a las glucoproteínas de la superficie celular resistencia frente a proteasas. Puesto que una de las características de Ecm33 es su riqueza en Ser/Thr y que forma parte de la superficie celular mediante el anclaje GPI, resulta de interés aportar datos de predicción de los sitios de glucosilación que presentamos, puesto que es probable esta modificación.

A la espera de que se determine la estructura proteica de algún miembro de la familia y con la cautela propia con la que en general deben de recibirse las predicciones de estructura secundaria, la primera característica en común desde un punto de vista estructural, es que se trata de un grupo de proteínas que adquiere una conformación principalmente *beta*, basándonos en el amplio consenso apreciado entre diferentes servidores de predicción de estructura.

En lo que respecta a los modelos de predicción de estructura recibidos a través de diferentes servidores seleccionados, existe consenso a la hora de establecer los parentales en la construcción de los modelos: la Internalina de *Listeria monocytogenes* o el Receptor del Factor de Crecimiento Insulina-like (IGFR). La elección de uno u otro parental está estrechamente relacionado con la parte de la secuencia que enviemos y el método empleado. Uno u otro parental son siempre seleccionados por los servidores de reconocimiento de plegamiento, con la excepción Robetta. Aunque Robetta emplea como parental en la construcción de sus *modelos comparativos* a la *Internalina* para el procesamiento de Ecm33, y también emplea un parental (IGFR) cuando enviamos la región N-terminal.



Ahora bien, si lo que enviamos es la región C-terminal (279-407) existe un cambio y Robetta procede al modelado *de novo*.

Pero analicemos las características de los parentales que son seleccionados por los métodos empleados. La Internalina (INLA\_LISMO) es una proteína que se encuentra en la superficie celular de *Listeria monocytogenes*. Esta proteína se encarga de interaccionar con una proteína humana, la Cadherina 1 (CAD1\_HUMAN), proteína que forma parte de la membrana de determinadas células humanas, de tal forma que la interacción entre ambas tiene como consecuencia la entrada en la célula humana de *L. monocytogenes*. Consultando la base de datos de clasificación de estructuras proteicas SCOP [14] sobre la región en concreto de esta proteína que sirve como parental para la elaboración de modelos de Ecm33, está catalogada de la clase alfa-beta, con un tipo de plegamiento de repeticiones ricas en Leucina (LRR), de la superfamilia de los L-Domains.

Con respecto a IGFR, se trata de un receptor que se dispone en la membrana plasmática de determinadas células humanas formando dímeros y que se caracteriza por la repetición de dos dominios (llamados L1 y L2) separados por una región rica en Cys. Estos dominios, llamados en general *L-Domain* son idénticos entre si y corresponde con la región del IGFR que es seleccionado como parental de cara a la elaboración de los modelos predictivos. SCOP clasifica los *L-domains* de la clase proteínas alfa-beta, con un tipo de pliegue de repeticiones ricas en leucina y de la superfamilia y familia de los *L-domains*.

Por tanto, se trata en realidad de un mismo tipo de conformación proteica la que es seleccionada de cara a la elaboración de los modelos estructurales, tanto en el caso del *fold recognition* como en el *modelado comparativo* de Robetta.

Aunque en principio todo hace pensar en función de los resultados obtenidos, que Ecm33 tuviera la conformación estructural completa que es dada en los modelos tanto de Robetta, 3D-PSSM y LOOPP, es conveniente subrayar que la parte terminal de la región C-terminal (del aminoácido 300 en adelante) presenta malas homologías estructurales en los diferentes métodos empleados, tanto para los modelos de la proteína al completo como para los resultados de la región c-terminal. De hecho, Robetta opta por el modelado *de novo*.

Con respecto a la exactitud de los modelos propuestos por Robetta empleando el protocolo *de novo* para esta región C-terminal, existen dos factores a tener en cuenta. Por un lado, que el método de modelado *de novo* está optimizado para fragmentos inferiores a 120 residuos. Pero por otro lado, si bien los mejores modelos obtenidos por Robetta en el LIVEBENCH 7 y 8 (<http://bioinfo.pl/LiveBench/>) [46], tanto para estructuras principalmente  $\alpha$  y principalmente  $\beta$  fueron para proteínas de entre 151-200 aminoácidos, en lo que respecta a las proteínas  $\alpha\beta$ , los resultados fueron mejores para dominios entre 100-150 aminoácidos [30]. Esto deja una puerta abierta para una posible buena predicción de los resultados de esta región C-terminal.

Con respecto a los resultados de las búsquedas que se obtienen mediante MAMMOTH de los modelos de la región c-terminal de Ecm33, el mejor Z-score obtenido corresponde al modelo 3 (Z-score = 6.32), que corresponde a una transferasa (Dihydropteroate synthase 1).

### **Sobre la función proteica**

En lo que respecta a la función que Ecm33 pudiera desempeñar en la superficie celular, es necesario detenernos en dos aspectos. Por un lado, el cambio que existe en la composición de la secuencia de Ecm33 hacia el tercio final de la secuencia. Si la primera parte era rica en repeticiones

aminoácidas y carga negativa, en esta región a la que hacemos referencia se ponen en común una mayor cantidad de aminoácidos que suelen desempeñar alguna función catalítica [47]. En este mismo sentido, el posible motivo estructural que identificamos en la posición 350 de la secuencia, carecería de valor significativo bajo el contexto de un modelado como en los propuestos por todos los servidores para la totalidad de la secuencia. Ahora bien, bajo el modelado *de novo* de esa región que nos ofrece Robetta, la región a la que hacemos referencia cobraría un mayor valor.

Si nos basamos en el paradigma de la secuencia-a-estructura-de estructura a función para predecir la función de esta proteína, a la vista del conjunto de modelos ofrecidos, estaríamos ante proteínas (la Familia de Ecm33) con una estructura característica de un receptor. El hecho de que tuviera que establecer una interacción con cualquier otro elemento podría verse reforzado por el patrón de carga que presenta la proteína (más negativa en el lado más N-terminal), región que además, según las predicciones, sufriría un mayor número de o-glucosilaciones. En este sentido, parece lógico pensar que interaccione con proteínas con función en la síntesis de la pared puesto que la delección del gen tiene como consecuencia una desorganización en la estructura de la pared.

En lo que respecta a su posible función catalítica, no puede afirmarse nada, pero si hubiera que señalar alguna región “responsable” de la acción catalítica, esta sería sin duda la región C-terminal. En este sentido y tras el análisis de los alineamientos múltiples, existe una característica que une a las dos Ecm33: si bien la cisterna 351 está muy conservada en todos los miembros de la familia, las dos Ecm33 se diferencian en esa región del resto de proteínas en la presencia de Aspártico en la región que precede al motivo de “hipotética” actividad.

Pero existe otra posibilidad si analizamos los datos de las bases de datos de interacciones proteicas. Ecm33 interacciona con varias proteínas que tienen algún tipo de función en el proceso de ubiquitinación o relacionadas con la maquinaria de degradación (*CDC53*, *ECM29*, *HTR1*, *RPN1*, *RPN8*, *RTT101*, *UBI4*). También interacciona con proteínas relacionadas con el tamaño celular y pared celular (*CDC39*, *HRT1*, *HYP2*), o bien proteínas relacionadas con el movimiento de orgánulos (*VPS13*, *UBI4*, *BBC1*). También existen proteínas relacionadas con respuestas frente a inanición frente a determinados aminoácidos (*GCN1*, *YGPI*) y otros tipos de stress (*TPS1*), y sin olvidar que interacciona con SLT2, una MAP kinasa relacionada principalmente con rutas de integridad de la pared celular. Teniendo en cuenta además los datos del estudio del *contexto genético*, SaEcm33 así como sus ortólogos (en este caso, CaEcm33), presentan genes con función en el transporte celular, la viabilidad celular, transporte de aminoácidos y sobre todo, relacionados con ubiquitinación (presentes en el contexto genético de todos los miembros de la familia Ecm33 en levadura).

Estos datos ayudan en el dibujo de otras posibilidades en cuanto al tipo de relaciones que establece esta proteína en la superficie celular. Si tenemos en cuenta la *desorganización* que se produce en la pared celular como consecuencia de la delección de Ecm33, más el hecho de que van disminuyendo los efectos de la desorganización a modo que vamos añadiendo *ecm33* (efecto fenotípico dosis-dependiente), es posible pensar que Ecm33 esté involucrada en el *turn-over* de la pared celular. Y esto podría ser a través de una participación activa en el orden de los procesos de construcción, a través del control de las proteínas implicadas mediante su relación con la maquinaria de degradación.

El mismo tipo de función desempeñarían en principio el resto de componentes de la Familia de Ecm33, cada uno en sus determinados contextos, tanto en la pared del organismo fúngico como en la pared de la espora. En función de la importancia de las relaciones entre cada uno de los

componentes de la familia y la maquinaria de degradación, tendría una mayor o menor importancia de cara al *turn-over* de la pared celular y sus efectos fenotípicos.

#### **Sugerencias experimentales**

- Mutar la “zona caliente” apreciada en la región c-terminal.
- Sobreexpresar determinados miembros de la maquinaria de degradación en mutantes *ecm33Δ* con el fin de ver si revierte dicha mutación.

**Tabla 1.** Predicción de Péptido Señal y GPI

	<b>Signal IP</b>	<b>Psort II</b>	<b>GPI</b>	<b>DGPI</b>	<b>GPI Dem</b>
<b>SaEcm33</b>	Si (21-22)	Si (19-20)	No (445)	SI (445)	Gpi: 407*
<b>CaEcm33</b>	Si (18-19)	Si (1-18)	Si (399)	Si (394)	
<b>SaPst1</b>	Si (19-20)	Si (24-25)	No (419)	No (419)	
<b>CaPst1</b>	Si (20-21)	Si (23-24)	No (384)	Si (385)	
<b>SaSps2</b>	No (23-24)	Si (56-57)	No (475)	No (475)	
<b>SaSps22</b>	Si (25-26)	Si (1-25)	Si (440)	Si (440)	
<b>CaSps22</b>	No (28-29)	No	Si (452)	No (440)	
<b>SpMeu10</b>	Si (29-30)	No (29-30)	No (395)	Si (389)	
<b>EgSps22</b>	Si (19-20)	Si (19-20)	No (457)	Si (455)	

\* [4]

**Tabla 2.** Tabla comparativa. En ella se recogen la composición aminoacídica de los principales miembros de la Familia de Ecm33, junto con los principales homólogos parentales seleccionados por los diferentes servidores de predicción de estructura.

<b>Ecm33 21-200</b>	<b>CaEcm33 19-204</b>	<b>SaPst1 19-199</b>	<b>CaPst1</b>	<b>SaSpS2</b>
Amino acid composition: Ala (A) 12 6.7% Arg I 1 0.6% Asn (N) 16 8.9% Asp (D) 11 6.1% Cys © 2 1.1% Gln (Q) 7 3.9% Glu (E) 6 3.3% Gly (G) 10 5.6% His (H) 0 0.0% Ile (I) 17 9.4% Leu (L) 18 10.0% Lys (K) 5 2.8% Met (M) 2 1.1% Phe (F) 8 4.4% Pro (P) 2 1.1% Ser (S) 35 19.4% Thr (T) 19 10.6% Trp (W) 0 0.0% Tyr (Y) 1 0.6% Val (V) 8 4.4%  - (Asp + Glu): 17 + (Arg + Lys): 6	Amino acid composition: Ala (A) 18 9.7% Arg I 0 0.0% Asn (N) 17 9.2% Asp (D) 10 5.4% Cys © 2 1.1% Gln (Q) 9 4.9% Glu (E) 8 4.3% Gly (G) 12 6.5% His (H) 1 0.5% Ile (I) 12 6.5% Leu (L) 18 9.7% Lys (K) 8 4.3% Met (M) 0 0.0% Phe (F) 7 3.8% Pro (P) 2 1.1% Ser (S) 17 9.2% Thr (T) 25 13.5% Trp (W) 0 0.0% Tyr (Y) 2 1.1% Val (V) 17 9.2%  - (Asp + Glu): 18 + (Arg + Lys): 8	Amino acid composition: Ala (A) 14 7.7% Arg I 1 0.6% Asn (N) 16 8.8% Asp (D) 9 5.0% Cys © 2 1.1% Gln (Q) 6 3.3% Glu (E) 3 1.7% Gly (G) 9 5.0% His (H) 1 0.6% Ile (I) 14 7.7% Leu (L) 19 10.5% Lys (K) 11 6.1% Met (M) 0 0.0% Phe (F) 8 4.4% Pro (P) 3 1.7% Ser (S) 35 19.3% Thr (T) 19 10.5% Trp (W) 0 0.0% Tyr (Y) 2 1.1% Val (V) 9 5.0%  - (Asp + Glu): 12 + (Arg + Lys): 12	Amino acid composition: Ala (A) 11 6.2% Arg I 0 0.0% Asn (N) 22 12.4% Asp (D) 10 5.6% Cys © 2 1.1% Gln (Q) 9 5.1% Glu (E) 5 2.8% Gly (G) 9 5.1% His (H) 0 0.0% Ile (I) 18 10.1% Leu (L) 24 13.5% Lys (K) 8 4.5% Met (M) 0 0.0% Phe (F) 6 3.4% Pro (P) 2 1.1% Ser (S) 26 14.6% Thr (T) 16 9.0% Trp (W) 0 0.0% Tyr (Y) 0 0.0% Val (V) 10 5.6%  - (Asp + Glu): 15 + (Arg + Lys): 8	Amino acid composition: Ala (A) 8 3.7% Arg I 4 1.8% Asn (N) 23 10.6% Asp (D) 12 5.5% Cys © 2 0.9% Gln (Q) 6 2.8% Glu (E) 17 7.8% Gly (G) 11 5.0% His (H) 4 1.8% Ile (I) 25 11.5% Leu (L) 24 11.0% Lys (K) 18 8.3% Met (M) 1 0.5% Phe (F) 8 3.7% Pro (P) 6 2.8% Ser (S) 16 7.3% Thr (T) 11 5.0% Trp (W) 2 0.9% Tyr (Y) 4 1.8% Val (V) 16 7.3%  - (Asp + Glu): 29 + (Arg + Lys): 22
<b>Sa SpS22</b>	<b>11gr (285-463)</b>	<b>1n8y (28-225)</b>	<b>1°6sAn</b>	
Amino acid composition: Ala (A) 8 3.6% Arg I 8 3.6% Asn (N) 23 10.4% Asp (D) 12 5.4% Cys © 3 1.4% Gln (Q) 11 5.0% Glu (E) 16 7.2% Gly (G) 9 4.1% His (H) 7 3.2% Ile (I) 24 10.9% Leu (L) 25 11.3% Lys (K) 12 5.4% Met (M) 1 0.5% Phe (F) 11 5.0% Pro (P) 9 4.1% Ser (S) 18 8.1% Thr (T) 8 3.6% Trp (W) 1 0.5% Tyr (Y) 2 0.9% Val (V) 13 5.9%  - (Asp + Glu): 28 + (Arg + Lys): 20	Amino acid composition: Ala (A) 6 3.1% Arg I 9 4.6% Asn (N) 15 7.7% Asp (D) 11 5.7% Cys © 7 3.6% Gln (Q) 9 4.6% Glu (E) 18 9.3% Gly (G) 14 7.2% His (H) 3 1.5% Ile (I) 13 6.7% Leu (L) 19 9.8% Lys (K) 14 7.2% Met (M) 5 2.6% Phe (F) 6 3.1% Pro (P) 4 2.1% Ser (S) 14 7.2% Thr (T) 9 4.6% Trp (W) 2 1.0% Tyr (Y) 6 3.1% Val (V) 10 5.2%  - (Asp + Glu): 29 + (Arg + Lys): 23	Amino acid composition: Ala (A) 8 5.9% Arg I 10 7.4% Asn (N) 8 5.9% Asp (D) 9 6.6% Cys © 2 1.5% Gln (Q) 14 10.3% Glu (E) 6 4.4% Gly (G) 9 6.6% His (H) 1 0.7% Ile (I) 6 4.4% Leu (L) 19 14.0% Lys (K) 5 3.7% Met (M) 2 1.5% Phe (F) 3 2.2% Pro (P) 7 5.1% Ser (S) 4 2.9% Thr (T) 5 3.7% Trp (W) 1 0.7% Tyr (Y) 4 2.9% Val (V) 13 9.6%  - (Asp + Glu): 15 + (Arg + Lys): 15	Amino acid composition: Ala (A) 14 5.0% Arg I 3 1.1% Asn (N) 39 13.9% Asp (D) 17 6.0% Cys © 0 0.0% Gln (Q) 11 3.9% Glu (E) 8 2.8% Gly (G) 9 3.2% His (H) 0 0.0% Ile (I) 23 8.2% Leu (L) 61 21.7% Lys (K) 12 4.3% Met (M) 1 0.4% Phe (F) 5 1.8% Pro (P) 9 3.2% Ser (S) 29 10.3% Thr (T) 29 10.3% Trp (W) 1 0.4% Tyr (Y) 3 1.1% Val (V) 7 2.5%  - (Asp + Glu): 25 + (Arg + Lys): 15	

**Tabla 3. Familia Ecm33.** Las proteínas identificadas han sido ordenadas por columnas y en función de la mayor homología con los miembros de *Saccharomyces cerevisiae*, única especie en la que se han conseguido verificar todas las proteínas de la familia de Ecm33.

<i>Saccharomyces cerevisiae</i> (Sc)	SaEcm33	SaPst1	SaSps2	SaSps22 YCE8
<i>Candida albicans</i> (Ca)	CaEcm33.3	CaPst1 Ecm33.1		CaSps22 IPF13972
<i>Schizosaccharomyces pombe</i> (Sp)	SpYin3	SpMeu10		
<i>Eremothecium gossypii</i> (Eg)	EgEcm33 Q75DT6			EgSps22 AFR723Cp
<i>Kluyveromyces lactis</i> (Kl)	01Kl gi 50305773			02Kl gi 50304567
<i>Neurospora crassa</i> (Nc)		NcMeu10 XP_3313281		
<i>Candida glabrata</i> (Cg)	01Cg gi 50294025 02Cg gi 50286919			03Cg gi 50288883 04Cg gi 50285137
<i>Debaryomyces hansenii</i> (Dh)	01Dh gi 50427223	03Dh ref XP_4625991		Dh gi 50428239 01Dh gi 50427223?
<i>Yarrowia lipolytica</i> (Yl)	01Yl gi 50550121			
<i>Aspergillus nidulans</i> (An)		An01 gi 40741117		
<i>Magnaporthe grisea</i> (Mg)		01Mg gb EAA530611		
<i>Gibberella zeae</i> (Gz)		01Gz gb EAA704051		

**Tabla 4.** Resultados de los diferentes servidores de Fold Recognition

Resultados de FFAS.						
Principales proteínas de la Familia Ecm33 y junto con fragmentos de SaEcm33						
Query	Length	Result vs.	Range	Score	%id	Covered by template(s)
SaEc33_pr	401	pdb0504	3-382	-25.200	12	1m6b_A mol:protein length:621 Receptor Protein-Tyrosine Kinase Erbb-3
			10-321	-25.700	13	1igr_A mol:protein length:478 Insulin-Like Growth Factor Receptor 1
SaEcm33_pr	401	scop165	6-165	-23.300	18	d1gra2 c.10.2.5 (A:300-478) Type 1 insulin-like growth factor receptor extrac-domain
			17-354	-15.800	11	d1o6va2 c.10.2.1 (A:33-416) Internalin A ( <i>L. monocytogenes</i> )
C-t SaEcm33	167	pdb0504	7-158	-19.600	12	1igr_A mol: 478 Insulin-Like Growth Factor Receptor 1
N-t SaEcm33	234	pdb0504	29-182	-28.900	17	1igr_A mol: 478 Insulin-Like Growth Factor Receptor 1
Ecm33p_Candida	423	scop165	33-130	-22.300	16	d1gra1 c.10.2.5 (A:1-149) Type 1 insulin-like growth factor receptor extrac-domain
			44-322	-12.500	14	d1o6va2 c.10.2.1 (A:33-416) Internalin A ( <i>Listeria monocytogenes</i> )
			59-285	-15.300	8	d1h6ua2 c.10.2.1 (A:36-262) Internalin H ( <i>Listeria monocytogenes</i> )
PST1_YEAST	444	pdb0504	27-320	-16.200	15	1m6b_A mol:protein length:621 Receptor Protein-Tyrosine Kinase Erbb-3
			29-396	-14.600	9	1igr_A mol:protein length:478 Insulin-Like Growth Factor Receptor 1
SPS2_YEAST	502	pdb0504	3-484	-15.200	9	1g9u_A mol:protein length:454 Outer Protein Yopm
			6-388	-26.100	13	1o6s_A mol:protein length:466 Internalin A
			101-443	-20.700	13	1igr_A mol:protein length:478 Insulin-Like Growth Factor Receptor 1
SPS22_Yeast	463	pdb0504	4-377	-16.700	12	1g9u_A mol:protein length:454 Outer Protein Yopm
			34-451	-12.800	7	1a4y_A mol:protein length:460 Ribonuclease Inhibitor
			36-415	-22.500	14	1o6s_A mol:protein length:466 Internalin A
			111-377	-25.000	13	1h6u_A mol:protein length:308 Internalin H

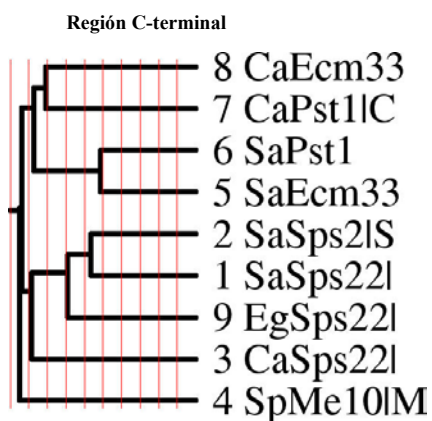
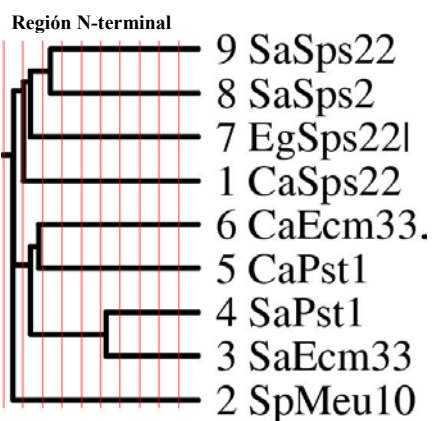
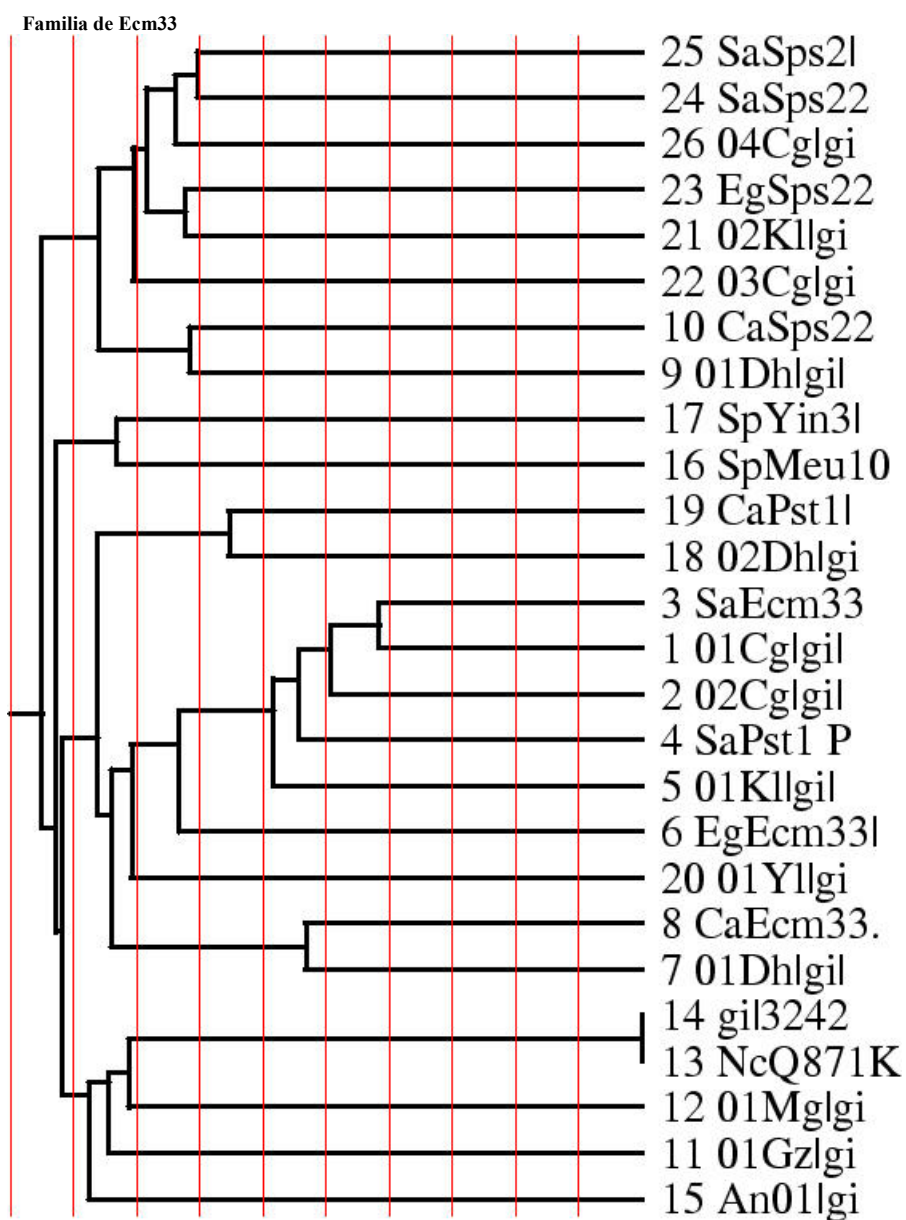
### Resultados de 3D-PSSM

SaEcm33 procesado (21-407)				
Fold Library	Template Length	PSSM E_Value	Fold	Superfamily
c1o6sa_15%i.d.	460	0.00154	bacterial infection	Internalin a
c1h6ua_14%i.d.	308	0.00809	Cell adhesion	Internalin h
c1j15a_13%i.d.	353	0.032	Toxin	outer protein yopm,.
c1h6ta_12%i.d.	291	0.0547	Cell adhesion.	Internalin b
Ecm240_407 C-terminal				
d1figra1_17%i.d.	149	3.84e-05	tyrosine,insulin,iii-like Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)	L domain-like
SaEcm33 (19-300) N-terminal				
c1h6ua_15%i.d.	308	0.133	membrane Signal Transmembrane	internalin

### Resultados de LOOPP para el procesamiento de Ecm33 (21-407)

Nombre	Confidence	score	Homolog	hits	length	Seq. Ident
IO6V_A	HIGH	3.142	0	1	97.94%	19.47%
IJL5_A	HIGH	1.901	0	1	85.05%	14.67%
IM9S_A	GOOD	1.763	0	1	97.94%	13.47%
IOZN_A	GOOD	1.632	1	1	68.30%	14.59%

**Figura 1.** Árboles filogenéticos mediante el empleo de TraceSuite 2.





**Figura 2.** Predicción de la estructura secundaria de SaEcm33

[illegible]

## Resultados de Robetta [30]

**Predicción 1.** Corresponde a la secuencia procesada de Ecm33p (19-407).

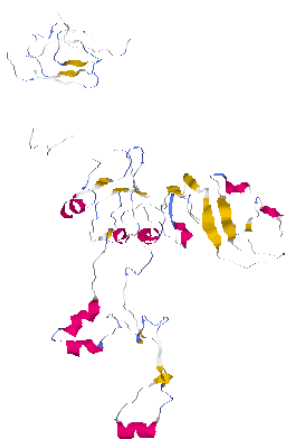
Es identificado un solo dominio a través de pdblast y el Pdb parental seleccionado es lo6sA correspondiente a la Internalina de *Listeria monocytogenes*. Mediante el método de alineamiento

Span	Source	Parent	Parent Span	Confidence	Annotations
1-388	pdblast	lo6sA_	81-356	8.698970	Bacterial Infection

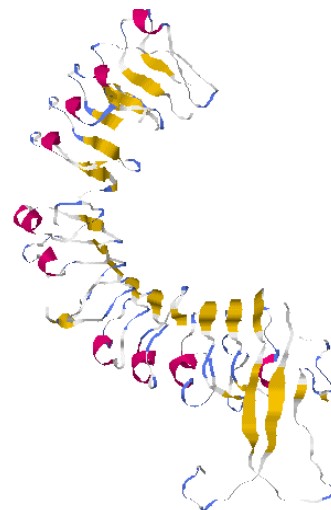
**Modelo1**



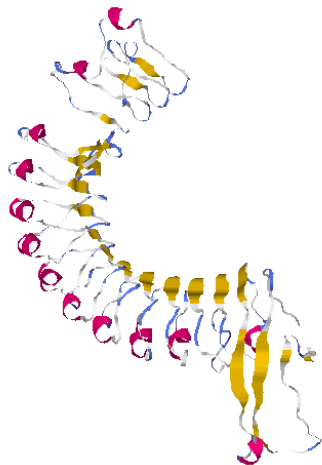
**Modelo 2**



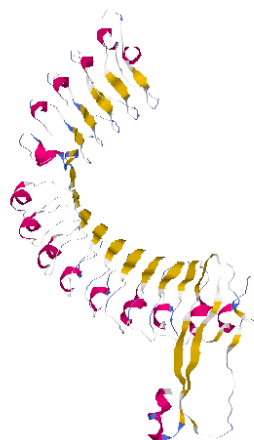
**Modelo3**



**Modelo 4**



**Modelo5**



**Predicción 2.** Correspondiente a los primeros 234 aa del extremo N-terminal de Ecm33.

1-234	pdblast	ligrA_	285-463	45.397940	Hormone Receptor
-------	---------	--------	---------	-----------	------------------

(Modelos aún en desarrollo)

**Predicción 3.** Corresponde a SaEcm33, el fragmento 279-407. El modelado en este caso fue *de novo*

279-407 cutpref denovo

cutpref -> de novo -> conf = 0 -> note: domain boundaries solely determined by sequence transitions, strongly predicted loop, occupancy, and distance from nearest block or terminus

Diez modelos son propuestos para este fragmento de la secuencia. Además de los modelos, se presentan los resultados de las búsquedas en MAMMOTH.

Los modelos están representados de tal forma que el extremo amino (aminoácido 279 de SaEcm33) está orientado hacia arriba. Las hélices alfa se representan en color cian, las láminas beta en amarillo, los giros en azul y el motivo Phosphatase-like en rojo.

**Modelo1**

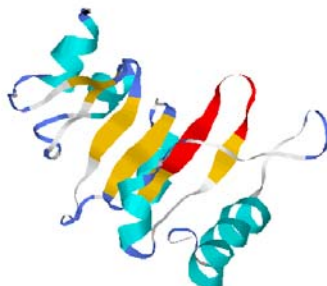
cb	co	description	env	hbbb	hs	pair	rama	rg	rsigma	score	sheet
44.3	19.95	F_0019_5206	-36.01	-46.2	-0.39	-35.61	97.01	15.52	-11.9	-101.23	0.69

**Modelo 2**

cb	co	description	env	hbbb	hs	pair	rama	rg	rsigma	score	sheet
46.26	19.35	F_0047_1338	-38.66	-50.44	0	-21.68	30.3	15.36	-16.21	-98.77	0.34
MAMMOTH Hits (Top 10 w/ minimum Z-score of 4.5)				Total Hits: 1							
Z-score	PDB ID	PDB Title		SCOP	Superfamily			Nsup	Nss	PDB resnum	
4.65	1mroC	METHANOGENESIS		d.58.31	Methyl-coenzyme M reductase subunits			127	67	247	

**Modelo 3**

cb	co	description	env	hbbb	hs	pair	rama	rg	rsigma	score	sheet
45.44	15.5	F_0046_0069	-36.65	-70.84	-2.49	-23.75	-21.63	16.1	-20.36	-115.22	0.34
MAMMOTH Hits (Top 10 w/ minimum Z-score of 4.5)				Total Hits: 33							
Z-score	PDB ID	PDB Title		SCOP Superfamily				Nsup	Nss	PDB resnum	
6.32	1eyeA	TRANSFERASE		c.1.21 Dihydropterolate synthetase-like				120	80	256	
6.09	1f6yA	TRANSFERASE		c.1.21 Dihydropterolate synthetase-like				128	75	258	
5.52	1qr6A	OXIDOREDUCTASE		c.58.1 Aminoacid dehydrogenase-like, N-terminal domain							
				c.2.1 NAD(P)-binding Rossmann-fold domains				127	81	537	
5.52	1euaA	LYASE		c.1.10 Aldolase				127	77	213	
5.39	1onrA	TRANSFERASE		c.1.10 Aldolase				128	88	316	
5.36	1rpxA	3-EPIMERASE		c.1.2 Ribulose-phosphate binding barrel				124	88	230	
5.34	1gox0	oxidoreductase (oxygen(a))		c.1.4 FMN-linked oxidoreductases				127	78	350	
5.31	1qopA	LYASE		c.1.2 Ribulose-phosphate binding barrel		123		88	265		
5.17	1ee2A	OXIDOREDUCTASE		b.35.1 GroES-like							
				c.2.1 NAD(P)-binding Rossmann-fold domains				127	85	373	
5.15	1aj20	SYNTHASE		c.1.21 Dihydropterolate synthetase-like				120	76	282	

**Modelo 4**

cb	co	description	env	hbbb	hs	pair	rama	rg	rsigma	score	sheet
46.54	17.91	F_0055_0446	-24.11	-53.7	-0.43	-22.9	-1.43	15.45	-21.8	-106.92	0.34
MAMMOTH Hits (Top 10 w/ minimum Z-score of 4.5)				Total Hits: 4							
Z-score	PDB ID	PDB Title		SCOP Superfamily				Nsup	Nss	PDB resnum	
5.02	1a800	OXIDOREDUCTASE		c.1.7 NAD(P)-linked oxidoreductase				78	277		
4.90	1eokA	HYDROLASE		c.1.8 (Trans)glycosidases				125	60	282	

4.87	1e79D	ATP PHOSPHORYLASE	a.69.1 C-terminal domain of alpha and beta subunits of F1 ATP synthase c.37.1 P-loop containing nucleoside triphosphate hydrolases									
4.50	1rpxA	3-EPIMERASE	b.49.1 N-terminal domain of alpha and beta subunits of F1 ATP synthase c.1.2 Ribulose-phosphate binding barrel	128 124	78 72	466 230						



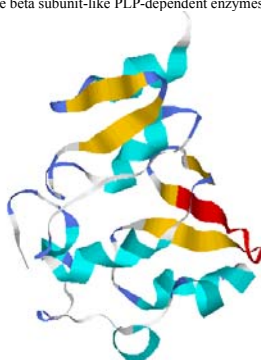
#### Modelo 5

cb	co	description	env	hbbb	hs	pair	rama	rg	rsigma	score	sheet
47.33	19.87	F_0016_0850	-30.27	-57.52	0.58	-33.02	9.91	16.32	-17.98	-101.96	2.99
MAMMOTH Hits (Top 10 w/ minimum Z-score of 4.5)											
Z-score	PDB ID	PDB Title		Total Hits: 1				Nsup	Nss	PDB resnum	
4.60	1bed0	OXIDOREDUCTASE		c.47.1	Thioredoxin-like		126	49	181		
				a.44.1	Disulphide-bond formation facilitator (DSBA), insertion domain						



#### Modelo 6

cb	co	description	env	hbbb	hs	pair	rama	rg	rsigma	score	sheet
44.62	20.38	F_0062_3232	-36.84	-68.39	-0.15	-32.71	17.61	14.79	-14.02	-100.91	0.34
MAMMOTH Hits (Top 10 w/ minimum Z-score of 4.5)											
Z-score	PDB ID	PDB Title		Total Hits: 1							
4.65	1fcjA	LYASE	c.79.1	Tryptophan synthase beta subunit-like PLP-dependent enzymes				127	86	302	



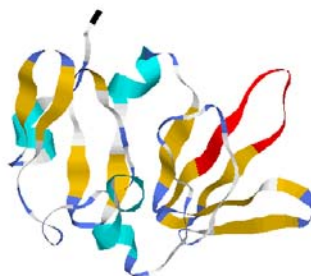
#### Modelo 7

cb	co	description	env	hbbb	hs	pair	rama	rg	rsigma	score	sheet
45.4	18.62	F_0017_7221	-39.36	-56.73	0.19	-28.91	27.61	15.41	-15.52	-103.94	1.57
MAMMOTH Hits (Top 10 w/ minimum Z-score of 4.5)											
Z-score	PDB ID	PDB Title		Total Hits: 2							
5.04	1iso0	OXIDOREDUCTASE		c.77.1	Isocitrate/Isopropylmalate dehydrogenases			128	70	414	
4.67	1iba0	phosphotransferase		d.95.1	Glucose permease domain IIB		75	30	78		



#### Modelo 8

cb	co	description	env	hbbb	hs	pair	rama	rg	rsigma	score	sheet
46.6	15.89	F_0054_0942	-31.87	-54.7	0.34	-30.15	10.26	16.1	-21.36	-106.78	0.69



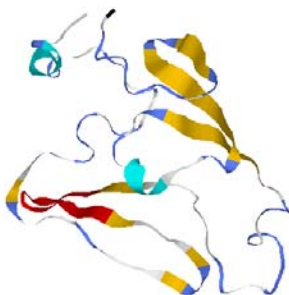
#### Modelo 9

cb	co	description	env	hbbb	hs	pair	rama	rg	rsigma	score	sheet
47.68	20.46	F_0100_6498	-44.27	-41.52	0.23	-37.25	-3.6	15.77	-15.09	-110.96	1.57
MAMMOTH Hits (Top 10 w/ minimum Z-score of 4.5) Total Hits: 1											
4.53	1tia0	hydrolase(carboxylic esterase)	c.69.1	alpha/beta-Hydrolases			121	76	271		

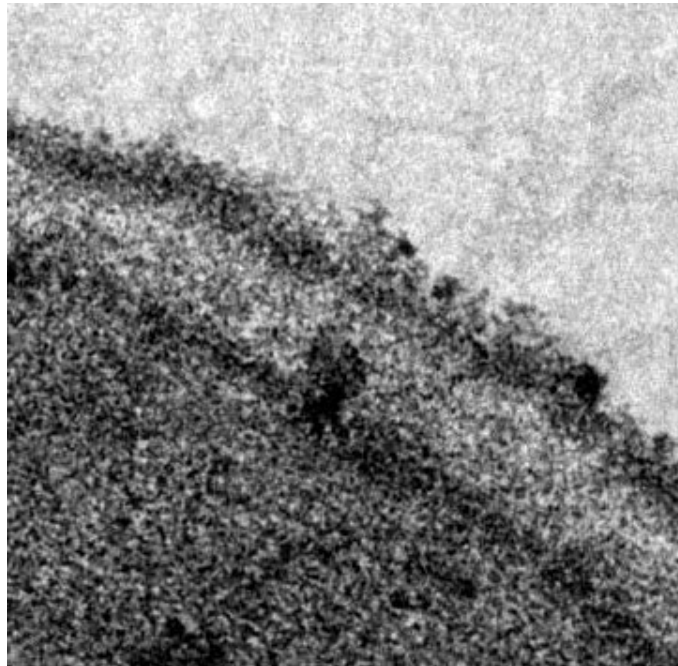


#### Modelo 10

cb	co	description	env	hbbb	hs	pair	rama	rg	rsigma	score	sheet
44.71	27.39	S_0011_7839	-42.37	-50.84	-0.31	-23.3	6.05	16.67	-26.81	-124.69	2.99
MAMMOTH Hits (Top 10 w/ minimum Z-score of 4.5) Total Hits: 1											
4.65	leyqA	ISOMERASE		d.36.1	Chalcone isomerase	127	80	212			



**Imagen 1.** Micrografía electrónica donde puede apreciarse el fenotipo del mutante *ecm33Δ* de *Saccharomyces cerevisiae* en la pared celular.



**Imagen 2.** Receptor de la Interleukina 6. En rojo el motivo en común con SaEcm33 (VCxxGxxST)



## Referencias

1. Martinez-Lopez R, Monteoliva L, Diez-Orejas R, Nombela C, Gil C: **The GPI-anchored protein CaEcm33p is required for cell wall integrity, morphogenesis and virulence in *Candida albicans*.** *Microbiology* 2004.Oct. 150,3341-3354.
2. Caro LH, Tettelin H, Vossen JH, Ram AF, van den EH, Klis FM: **In silicio identification of glycosyl-phosphatidylinositol-anchored plasma-membrane and cell wall proteins of *Saccharomyces cerevisiae*.** *Yeast* 1997.Dec. 1913,1477-1489.
3. de Groot PW, Hellingwerf KJ, Klis FM: **Genome-wide identification of fungal GPI proteins.** *Yeast* 2003.Jul.15. 1920,781-796.
4. Terashima H, Hamada K, Kitada K: **The localization change of Ybr078w/Ecm33, a yeast GPI-associated protein, from the plasma membrane to the cell wall, affecting the cellular function.** *FEMS Microbiol.Lett.*2003.Jan.21 218,175-180.
5. Lussier M, White AM, Sheraton J, di Paolo T, Treadwell J, Southard SB, Horenstein CI, Chen-Weiner J, Ram AF, Kapteyn JC, Roemer TW, Vo DH, Bondoc DC, Hall J, Zhong WW, Sdicu AM, Davies J, Klis FM, Robbins PW, Bussey H: **Large scale identification of genes involved in cell surface biosynthesis and architecture in *Saccharomyces cerevisiae*.** *Genetics* 1997.Oct. 147,435-450.
6. Oguchi T, Oguchi T: **Genetic characterization of genes encoding enzymes catalyzing addition of phospho-ethanolamine to the glycosylphosphatidylinositol anchor in *Saccharomyces cerevisiae*.** *Genes Genet.Syst.*2002.Oct. 1977,309-322.
7. Higgins VJ, Alic N, Thorpe GW, Breitenbach M, Larsson V, Dawes IW: **Phenotypic analysis of gene deletant strains for sensitivity to oxidative stress.** *Yeast* 2002.Feb. 1919,203-214.
8. Garcia R, Bermejo C, Grau C, Perez R, Rodriguez-Pena JM, Francois J, Nombela C, Arroyo J: **The global transcriptional response to transient cell wall damage in *Saccharomyces cerevisiae* and its regulation by the cell integrity signaling pathway.** *J.Biol.Chem.*2004.Apr 9. 279,15183-15195.
9. Jung US, Levin DE: **Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway.** *Mol.Microbiol.*1999.Dec. 1934,1049-1057.
10. Percival-Smith A, Segall J: **Characterization and mutational analysis of a cluster of three genes expressed preferentially during sporulation of *Saccharomyces cerevisiae*.** *Mol.Cell Biol.* 1986, 6:2443-2451.
11. Tohe A, Oguchi T: **Las21 participates in extracellular/cell surface phenomena in *Saccharomyces cerevisiae*.** *Genes Genet.Syst.*1999.Oct. 1974,241-256.
12. Fetrow JS, Skolnick J: **Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases.** *J.Mol.Biol.*1998.Sep.4 281,949-968.
13. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective .** *J.Mol.Biol.*2001.Apr 6. 307,1113-1143.
14. Lo CL, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C: **SCOP: a structural classification of proteins database.** *Nucleic Acids Res.*2000.Jan.1 1928,257-259.
15. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR: **The Pfam protein families database.** *Nucleic Acids Res.*2004.Jan.1 1932,D138-D141.
16. Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM: **Protein folds and functions.** *Structure.*1998.Jul.15. 1906,875-884.

17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.*1997.Sep.1 1925,3389-3402.
18. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc.Natl.Acad.Sci.U.S.A* 1988.Apr 1985,2444-2448.
19. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment** . *J.Mol.Biol.*2000.Sep.8. 302,205-217.
20. Eddy SR: **Profile hidden Markov models.** *Bioinformatics.*1998. 1914,755-763.
21. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J.Mol.Biol.*1996.Mar.29. 257,342-358.
22. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J.Mol.Biol.*2004.Jul.16. 340,783-795.
23. Eisenhaber B, Bork P, Eisenhaber F: **Prediction of potential GPI-modification sites in proprotein sequences.** *J.Mol.Biol.*1999.Sep.24. 292,741-758.
24. Rost B: **PHD: predicting one-dimensional protein structure by profile-based neural networks.** *Methods Enzymol.*1996. 266,525-539.
25. Rost B, Yachdav G, Liu J: **The PredictProtein server.** *Nucleic Acids Res.*2004.Jul.1 1932,W321-W326.
26. Jaroszewski L, Rychlewski L, Godzik A: **Improving the quality of twilight-zone alignments.** *Protein Sci.*2000.Aug. 1909,1487-1496.
27. Rychlewski L, Jaroszewski L, Li W, Godzik A: **Comparison of sequence profiles. Strategies for structural predictions using sequence information.** *Protein Sci.*2000.Feb. 1909,232-241.
28. Kelley LA, MacCallum RM, Sternberg MJ: **Enhanced genome annotation using structural profiles in the program 3D-PSSM.** *J.Mol.Biol.*2000.Jun.2 299,499-520.
29. Teodorescu O, Galor T, Pillardy J, Elber R: **Enriching the sequence substitution matrix by structural information.** *Proteins* 2004.Jan.1 1954,41-48.
30. Kim DE, Chivian D, Baker D: **Protein structure prediction and analysis using the Robetta server.** *Nucleic Acids Res.*2004.Jul.1 1932,W526-W531.
31. Rohl CA, Strauss CE, Chivian D, Baker D: **Modeling structurally variable regions in homologous proteins with rosetta.** *Proteins* 2004.May.15. 1955,656-677.
32. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D: **Automated prediction of CASP-5 structures using the Robetta server.** *Proteins* 2003. 1953,524-533.
33. Ortiz AR, Strauss CE, Olmea O: **MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.** *Protein Sci.*2002.Nov. 1911,2606-2621.
34. Sayle RA, Milner-White EJ: **RASMOL: biomolecular graphics for all.** *Trends Biochem.Sci.*1995.Sep. 1920,374.
35. Vazquez A, Flammini A, Maritan A, Vespignani A: **Global protein function prediction from protein-protein interaction networks.** *Nat.Biotechnol.*2003.Jun. 1921,697-700.
36. Deng M, Zhang K, Mehta S, Chen T, Sun F: **Prediction of protein function using protein-protein interaction data** . *J.Comput.Biol.*2003. 1910,947-960.



37. Bader GD, Hogue CW: **Analyzing yeast protein-protein interaction data obtained from different sources.** *Nat.Biotechnol.*2002.Oct. 1920,991-997.
38. Ostanin K, Pokalsky C, Wang S, Van Etten RL: **Cloning and characterization of a *Saccharomyces cerevisiae* gene encoding the low molecular weight protein-tyrosine phosphatase.** *J.Biol.Chem.*1995.Aug.4 270,18491-18499.
39. Yawata H, Yasukawa K, Natsuka S, Murakami M, Yamasaki K, Hibi M, Taga T, Kishimoto T: **Structure-function analysis of human IL-6 receptor: dissociation of amino acid residues required for IL-6-binding and for IL-6 signal transduction through gp130.** *EMBO J.*1993.Apr 1912,1705-1712.
40. Schwede T, Kopp J, Guex N, Peitsch MC: **SWISS-MODEL: An automated protein homology-modeling server.** *Nucleic Acids Res.*2003.Jul.1 1931,3381-3385.
41. Lund O, Frimand K, Gorodkin J, Bohr H, Bohr J, Hansen J, Brunak S: **Protein distance constraints predicted by neural networks and probability density functions.** *Protein Eng* 1997.Nov. 1910,1241-1248.
42. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ: **Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM.** *Proteins* 2001. 2005,39-46.
43. Lambert C, Leonard N, De B, X, Depiereux E: **ESyPred3D: Prediction of proteins 3D structures.** *Bioinformatics.*2002.Sep. 1918,1250-1256.
44. de Groot PW, de Boer AD, Cunningham J, Dekker HL, de Jong L, Hellingwerf KJ, de Koster C, Klis FM: **Proteomic analysis of *Candida albicans* cell walls reveals covalently bound carbohydrate-active enzymes and adhesins.** *Eukaryot.Cell* 2004.Aug. 2003,955-965.
45. Van den SP, Rudd PM, Dwek RA, Opdenakker G: **Concepts and principles of O-linked glycosylation.** *Crit Rev.Biochem.Mol.Biol.*1998. 1933,151-208.
46. Rychlewski L, Fischer D, Elofsson A: **LiveBench-6: large-scale automated evaluation of protein structure prediction servers.** *Proteins* 2003. 1953,542-547.
47. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM: **Analysis of catalytic residues in enzyme active sites.** *J.Mol.Biol.*2002.Nov.15. 324,105-121.