UNIVERSIDAD AUTÓNOMA METROPOLITANA UNIDAD XOCHIMILCO

Casa abierta al tiempo

DIVISIÓN DE CIENCIAS BIOLÓGICAS Y DE LA SALUD

DEPARTAMENTO DE SISTEMAS BIOLÓGICOS

LICENCIATURA EN QUÍMICA FARMACÉUTICA BIOLÓGICA
INFORME DE ACTIVIDADES DEL SERVICIO SOCIAL:

Caracterización *in silico* de la vía de señalización de respuesta a estrés osmótico en hongos filamentosos de la división *Ascomycota*.

PROYECTO GENÉRICO:

Obtención de materias primas, principios activos, medicamentos y productos biológicos.

PRESENTA:

Zarza Sánchez Larissa

Matrícula: 2173082232

Tutores:

Dr. Juan Esteban Barranco Florido No. Eco. 24927

Dr. Jesús Eduardo Zúñiga León Cédula Profesional 10977107

LUGAR DE REALIZACIÓN: Laboratorio De Biotecnología, Edificio N. Departamento de Sistemas Biológicos, Universidad Autónoma Metropolitana Unidad Xochimilco, con Dirección: Calzada Del Hueso 1100. Col. Villa Quietud, Alcaldía de Coyoacán, C.P 04960, Ciudad de México, México.

Periodo: 17 de noviembre de 2021 al 17 de mayo de 2022

Septiembre 2023

Tabla de Contenido

1. INTRODUCCIÓN

A lo largo de los años los hongos han sido un importante tema de estudio científico por sus peculiares características metabólicas, que a través de la investigación y experimentación se han logrado obtener interesantes sustancias de beneficio químico – farmacéutico a niveles industriales, por lo que estos han sido aprovechados ampliamente.

Los hongos adoptan diversas estructuras morfológicas diferenciadas dependientes de condición; en tales circunstancias los mohos son organismos multicelulares que forman redes de hifas, las cuales se encargan de dar estructura a los hongos filamentosos y de captar los nutrientes, además de ser altamente adaptativos a diversos ambientes dada su respuesta a diferentes tipos de estrés.

De tal modo, uno de los grupos de hongos de interés, son aquellos que pertenecen a la división *Ascomycota*, los cuales cuentan con una gran variedad de géneros que secretan una amplia diversidad de enzimas, proteínas y metabolitos secundarios de interés industrial.

La bioinformática ha resultado una herramienta muy útil para organizar, identificar, comparar y analizar la diferente información biológica como proteínas o metabolitos de interés en ascomicetos, que por medio de los estudios filogenéticos se han logrado identificar relaciones taxonómicas partiendo de secuencias.

Gracias a la proteómica que estudia la estructura y función de las proteínas que conforman el proteoma, y a las distintas bases de datos en las cuales se interpreta y analiza los datos, se puede estudiar más fácilmente la funcionalidad de cualquier grupo de proteínas y los mecanismos celulares y vías de señalización en las que se encuentran implicadas. En el presente trabajo se pretende identificar proteínas ortólogas de la vía de señalización de respuesta a estrés osmótico en hongos filamentosos a partir proteínas caracterizadas de *Aspergillus nidulans* un hongo modelo *Ascomycota,* y así comparar la dinámica molecular de esta vía. Este resultado también nos permitirá determinar si se conserva la misma vía de señalización en todas las especies de hongos filamentosos estudiados o si existen proteínas exclusivas de género o especie.

2. Marco Teórico

2.1 Características generales de hongos filamentosos

Los hongos son unicelulares o multicelulares y, dependiendo del organismo, pueden tomar múltiples estructuras morfológicas basadas en el modo de división y crecimiento celular. En general, los hongos crecen como levaduras u hongos filamentosos, los cuales también son llamados mohos (McGinnis, MR., Tyring SK., 1996).

Los hongos filamentosos son multicelulares ya que crecen por extensión apical de sus filamentos, conocidas como hifas (**Figura 1**). Este crecimiento hifal puede ocurrir con o sin separación de la pared celular de los compartimentos celulares, lo que se conoce como septación. El crecimiento continuo en los puntos de germinación da como resultado una compleja red de hifas conocida como micelio, el cual se divide en el micelio vegetativo que capta los nutrientes y humedad necesaria para el moho y el micelio aéreo conformado por hifas aéreas que portan las estructuras reproductivas (esporas teleomorfas). Si estas hifas se ramifican mientras crecen o no, y las estructuras macroscópicas de esta ramificación si se producen, es una de las muchas características utilizadas para la diferenciación de grupos de hongos. (Cole G. T.,1996; Ebbole D. J., 1996).

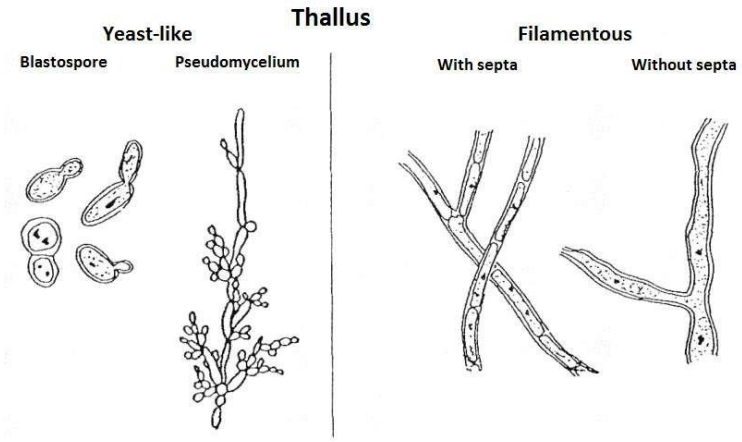

Figura 1. Hongos tipo levadura y filamentosos.

Las levaduras son formas unicelulares que se reproducen por gemación, mientras que los mohos forman hifas multicelulares (Kendrick B., 1985).

2.2 Clasificación taxonómica

Este grupo de hongos saprofitos con morfología filamentosa se localiza en la subdivisión *Pezizomycota,* la cual pertenece a la división *Ascomycota* (Hibbett *et al*., 2007), en la que encontramos una enorme diversidad de géneros que se han vuelto de gran interés por la comunidad científica debido a que sus procesos metabólicos se pueden utilizar para producir y refinar una amplia gama de productos (Hüttner *et al.,* 2020) como agentes de control biológico, proteínas y enzimas, productos farmacéuticos (antibióticos, inmunoestimulantes, inmunosupresores, estatinas, entre otros), productos químicos (ácidos orgánicos), etc. (Egbuta *et al.,* 2016). No obstante, la división Ascomycota incluye hongos patógenos de plantas (fitopatógenos), insectos (entomopatógenos) y humanos, que, en consecuencia, afectan o benefician de forma significativa a diversos sectores (economía, farmacéutico, salud, agrícola, silvicultura, entre otros). Por lo que es importante su óptima identificación y diferenciación más allá de su morfología. Los estudios filogenéticos han ayudado a identificar relaciones evolutivas y estimar la similitud entre organismos (**Figura 2**).

Dothideomycetes, Eurotiomycetes, Leotiomycetes y Sordariomycetes han sido hasta ahora las clases con estudios filogenómicos a gran escala con respecto a la división *Ascomycota*, debido a que presentan géneros que son de gran importancia en la actualidad (Ver más en el sig. apartado) (Muggia *et al.,* 2020). Apoyándose de las herramientas bioinformáticas y bases de datos como (UniProt, NCBI, InterPro, Pfam EMBOSS, Jalview, Clustal Omega, iTool, etc.) y lenguajes de programación como Python, Perl, R y Java, que han permitido facilitar el análisis de una amplia variedad de proteínas (Persson B., 2000; Mangalam H., 2002; Pazos, F., & Chagoyen M., 2015). Ver más adelante.

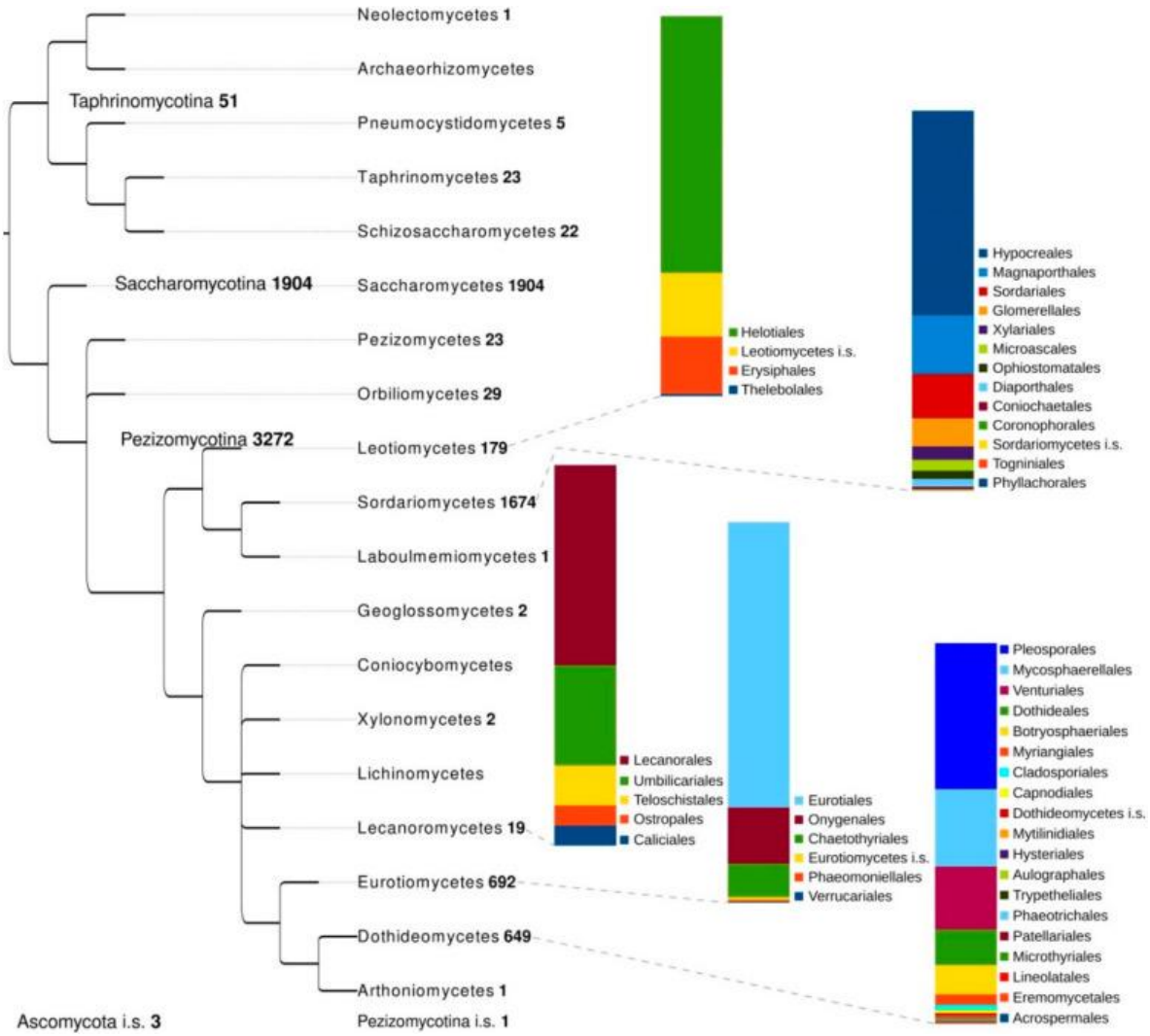

Figura 2. *Ascomycota* el árbol de la vida.

La topología, los subfilos y las clases incluidas se toman de Spatafora *et al.* (2017). Junto a cada taxón se informa el número de ensamblajes recuperados de la base de datos del NCBI (6 de julio de 2020). Para las cinco clases más diversas de *Ascomycota* también se reportan órdenes que tienen al menos un genoma secuenciado; los gráficos de barras están en porcentaje. Los datos se recuperaron de los informes de ensamblaje de NCBI, utilizando el script ncbi-genome-download (https://github.com/kblin/ncbi-genome-download/) Apache License (Version 2.0, January 2004) y un script personalizado que explota las utilidades de programación NCBI Entrez para recuperar la taxonomía (Muggia *et al*., 2020).

2.3 La importancia actual de los hongos filamentosos

Los hongos filamentosos son conocidos por ser saprófitos, ya que antes de asimilar nutrientes, ellos degradan la materia a compuestos más simples, por ejemplo; los polisacáridos vegetales (celulosa, almidón, pectina) a monosacáridos (glucosa, oligosacáridos), dada su gran y diversa secreción de enzimas (cuentan con 100-250 genes que codifican para enzimas biocatalíticas) y su eficiente capacidad hidrolítica de la biomasa vegetal (Spatafora *et al.*, 2017). Debido a ello, se han considerado como un potencial biotecnológico para ser explotados industrialmente, siendo catalogados como “Fabricas de células productoras de enzimas y proteínas de alto rendimiento” (Lübeck, M. & Lübeck, P. S., 2022).

Aunado a esto, a lo largo de los años, la fabricación de múltiples productos derivados de los hongos ha sido optimizada por la ingeniería genética, con el fin de obtener una sobreproducción de algún metabolito de interés, mejorar la expresión y excreción de proteínas, disminuir los costes de producción y cumplir con la demanda de industrialización (ante la sobrepoblación, así como los altos estándares de calidad), además de obtener sostenibilidad de las células fúngicas como recurso de base biológica para una economía circular sustentable (Meyer *et al.*, 2016; 2020), la cual se basa en la producción, el consumo y la recuperación (Interreg, 2020). En un futuro se espera, sea competitiva con la industria petroquímica, ya que ahora tiene limitaciones (Calisto *et al.*, 2020).

Algunas de las herramientas de la ingeniería genética para obtener estas mejoras son:

- Secuenciación de proteomas completos.
- Exploración completa de los sistemas fúngicos para la óptima manipulación simultanea de múltiples vías.
- El diseño de nuevas proteasas para aumentar el rendimiento de la cepa fúngica.
- La edición del genoma de cepas fúngicas a través de la recombinación de ADN, RNAi y la tecnología CRISPR-Cas (Wang *et al.,* 2020).

Empresas como AB Enzymes, BASF, Bayer, Chr. Hansen, DSM, DuPont, Novozymes, Puratos y Roal Oy, actualmente son lideres internacionales en el uso de estos hongos para la fabricación a granel de ácidos orgánicos, proteínas, enzimas y metabolitos secundarios (Meyer *et al*, 2016; 2020).

Así estos productos de origen fúngico benefician en gran medida a múltiples sectores, aunque aún existen ciertas limitaciones:

El género *Aspergillus* ha sido ampliamente explotado industrialmente para la producción de enzimas (Mojsov, K. D., 2016; Cairns *et al.*, 2018) ácidos orgánicos (Yang *et al.*, 2017) y metabolitos secundarios (Soltani, 2016), gracias a ello se ha mantenido en auge su extensa investigación biotecnológica. Así es como *A. nidulans/Emericella nidulans* (anamorfo) se le ha establecido como un organismo huésped para la manipulación genética con el fin de mejorar la síntesis o secreción de enzimas heterólogas como: lipasas (ampliamente utilizadas en la industria oleoquímica, manufactura de detergentes y en la industria alimenticia), xilanasas (útil en la industria de papel y pulpa, para el blanqueamiento de pulpa kraft y bambú), entre otras, por lo que se le ha catalogado como uno de los representantes de *Aspergillus* (Meyer *et al.*, 2016; Martínez *et al.,* 2019; Kumar, 2020).

Acremonium chrysogenum por otro lado es el principal productor industrial del antibiótico betalactámico cefalosporina C, del cual se derivan las demás cefalosporinas y que actualmente son de las más utilizadas en la práctica clínica (dada la prevalencia de enfermedades infecciosas y el aumento de financiamiento para el desarrollo de antibióticos/optimizar su producción a través de la biotecnología). Resultado de ello se estima que su mercado a nivel mundial incremente de 13 mil millones (2019) a 16 mil millones (2027) (Surabhi, P. & Onkar, S., 2020; Liu *et al*., 2022).

Otro hongo con gran potencial biotecnológico es el ascomiceto termófilo *Myceliophthora thermophila*, el cual tiene un crecimiento óptimo a 45°C y secreta la mayoría de las enzimas hidrolíticas (celulasas, lacasas, xilanasas, pectinasas, lipasas, fitasas, entre otras), siendo estas, termoestables incluso a altas temperaturas (70 - 80°C), estas degradan más rápido la biomasa vegetal (celulosa, lignocelulósica, hemicelulosa y otros) que las enzimas de hongos mesófilos como *Trichoderma* o *Aspergillus*, a altas temperaturas (Singh, 2014). Algunas aplicaciones son: producción de biocombustibles a partir de lignocelulosa (proceso no contaminante, aunque con ciertas limitaciones) (Raud *et al.*, 2019), biomasa como materia prima en la industria de pulpa y papel, reciente producción industrial de enzimas recombinantes (fitasas, lacasas, β-manosidasa, feruloil esterasa) con huéspedes fúngicos (*Aspergillus oryzae, A. niger, Pichia pastoris)* (Li, J. *et al.*, 2020).

A su par el género *Fusarium* tiene un futuro prometedor en la biotecnología industrial debido a la producción de enzimas hidrolíticas, metabolitos secundarios activos, pigmentos, aromas y biocombustibles como él biofuel (a través de enzimas como endoglucanasa y xilanasa para la hidrolisis de la biomasa) (Pessôa *et al.*, 2017). Sin embargo, producen micotoxinas (tricotecenos, zearalenonas y fumonisinas) que perjudican año tras año al sector de la

agricultura, dañando cultivos y haciéndolos inadecuados para el consumo. Es el caso de *Fusarium graminearum (Gibberella zeae),* nombrado Fusarium Head Blight (FHB) dada su gran patogenicidad en cultivos de trigo de Europa, América del Norte y Asia, en consecuencia, ha causado pérdidas millonarias (USD 1.176 mil millones en 2015-2016 en E.U.) (Wilson *et al.*, 2018), y se ha reportado la pérdida de 3.20% y 8.75% de rendimiento en el cultivo de trigo en zonas de EU/Canadá y China respectivamente (Savary *et al.*, 2019). Debido a ello y con el fin de contrarrestar/eliminar esta problemática se realizó un estudio *in silico* sobre la conservación del transcriptoma en cepas agresivas, en el que se identificó que compartían un 90% los patrones de expresión durante la infección, lo que demuestra su gran conservación y abre puertas a posibles soluciones (Rocher *et al.*, 2022).

Otros hongos fitopatógenos son: *Botryotinia fuckeliana* o también llamada *Botrytis cinérea*, que afecta a más de 200 especies de plantas en el mundo, principalmente las dicotiledóneas (rosal, menta, habas) y otros cultivos importantes como el garbanzo, frijol, brócoli, las frambuesas, moras, uvas y fresas, causando la enfermedad de moho gris (Cheung *et al.,* 2020).

Zymoseptoria tritici causa tizón foliar o Septororia tritici blotch (STB) en el trigo duro, es muy invasivo en zonas templadas o en épocas de lluvias. En Europa el 70% del uso anual de fungicidas para el trigo, está relacionado con STB, aunado a esto, se han reportado pérdidas de hasta el 20% de la cosecha anual, considerando que la pérdida anual oscila entre 5-10 % y los costos directos son de 120-700 millones de euros, este patógeno representa un gran problema en el sector agrícola (Fones, H., & Gurr, S., 2015).

Blumeria graminis mildiú polvoroso también se encuentra entre las enfermedades más importantes de los cereales en la actualidad (Różewicz *et al*., 2021) afectando a regiones de todo el mundo: Europa, América del Norte y Sur, Asia central, Egipto, Sudáfrica y China.

Colletotrichum gloeosporioides o el hongo de la antracnosis; provoca la pudrición de un amplio rango de hospedantes (Rojo *et al.,* 2017) como los cultivos de frutas tropicales: mango (Huerta *et al*., 2009), aguacate (Tapia *et al.*, 2020), papaya (Santamaría *et al*., 2011; Molina *et al*., 2017), manzana, piña y sandía, así como cereales, legumbres y verduras en países productores como México, Chile, Guatemala, Colombia, República dominicana, Perú, Costa Rica, Brasil, Estados Unidos, India, Indonesia, Pakistán y China, afectando su economía desde sus primeros reportes en 1903, hasta la actualidad (Sharma, M. & Kulshrestha, S., 2015).

Entre las alternativas contra dichas enfermedades causadas por estos fitopatógenos se encuentran los llamados hongos biofungicidas o agentes de biocontrol (BCA), los cuales a través de mecanismos como el micoparasitismo, la competición de espacio y nutrientes, la

secreción de enzimas hidrolíticas y la antibiosis/secreción de metabolitos, reducen la cantidad y/o efecto de los fitopatógenos previniendo la aparición de la enfermedad o su esparcimiento. Además, son capaces de formar un talo fúngico que beneficia a la planta al facilitar el acceso y la absorción de nutrientes, así como mejorar su resistencia (Thambugala *et al.,* 2020).

Los géneros más conocidos como hongos biocontrol son: *Trichoderma, Aspergillus, Penicillium, Pichia, Alternia*, entre otros. *Trichoderma* es el género más reconocido ya que engloba a 25 especies BCA, todas ellas tienen alta capacidad para sobrevivir en condiciones desfavorables, al usar los nutrientes a su conveniencia, también poseen una alta capacidad reproductiva, capacidad de modificar la rizosfera, fuerte agresividad contra hongos fitopatógenos y promueve el crecimiento vegetal/mecanismos de defensa. Las especies más relevantes son: *Trichoderma reesei, Trichoderma virens, Trichoderma atroviride, Trichoderma harzianum* (se ha comprobado que algunas cepas modifican el pH externo a conveniencia según la secreción de sus enzimas) y *Trichoderma asperellum* (Benítez *et al.,* 2004; Adnan *et al.,* 2019).

Algunas otras aplicaciones de interés reciente son: la producción de queratinasa por *T. harzianum* a partir de plumas de pollo (considerado un desecho), con el fin de ser aplicada en industrias de cuero y detergentes (Bagewadi *et al.,* 2018). La degradación de plástico PET y poliuretano a escala laboratorio por *Acremonium, Alternaria, Aspergillus, Emericella, Fusarium, Penicillum, Trichoderma,* entre otros como alternativa ante la contaminación de plástico. (Pathak, V. M., & Navneet, 2017; Meyer *et al.*, 2020). La **Figura 3** engloba las diversas industrias que se benefician de los hongos, dentro de los cuales, los *Ascomycota* son realmente llamativos dadas sus características ya descritas.

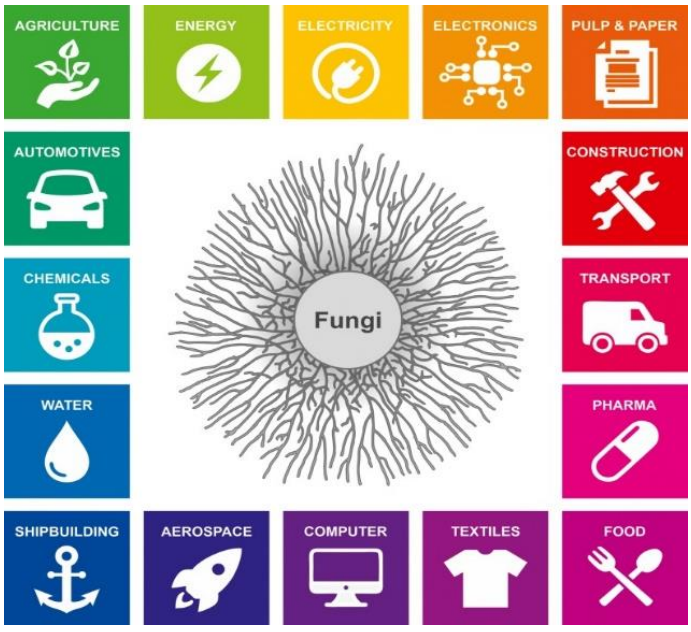

Figura 3. Industrias que se benefician de las capacidades metabólicas de los hongos (Meyer *et al.,* 2020).

Aunque se han encontrado múltiples aplicaciones debemos saber que también existen hongos peligrosos para los humanos, que ocasionan infecciones y enfermedades, y que en la actualidad se han visto en aumento por el envejecimiento de la población y/o el crecimiento del número de personas inmunodeprimidas por diversas causas (quimioterapias, trasplante de órganos, enfermedades como SIDA, Covid-19, diabetes mellitus, etc) (Rokas, 2022). Estos se han encontrado en diferentes géneros del reino fúngico/*Ascomycota*:

En *Sporothrix,* las especies *schenckii, globosa y brasiliensis* provocan la esporotricosis causando lesiones en la piel a diferentes niveles, así como a mucosas y a nivel sistémico (Barros *et al.,* 2011). *Sporothrix schenckii* se ha reportado como el agente etiológico más prevalente en las Américas en los últimos 10 años. Las principales zonas endémicas de *S. schenckii* en el mundo son: Australia, Sudáfrica, América del Norte y América del Sur (Hernández *et al*., 2022).

Trichophyton rubrum es otro hongo dermatofito, que en particular produce tinea capitis (pelo), tinea corporis (corporal), tinea pedis (pie de atleta), tinea manuum (mano) y tinea unguium u onicomicosis (uñas), y que prevalece en muchas ocasiones después de un tratamiento con antifúngicos (INSST, 2021). Estas infecciones superficiales afectan alrededor del 25% de la población mundial (1.7 mil millones) y aumenta al 50% en personas de 70 años o más (Brown *et al*., 2012).

Pneumocystis jirovecii es el agente causal de la neumonía en humanos, reportando >400.000 casos al año con una tasa de mortalidad del 20-80%, la cual aumenta en personas inmunodeprimidas, personas mayores y niños con VIH (Brown *et al.,* 2012; Meyer *et al*., 2016).

Así mismo *Lomentospora prolificans* que causa lomentosporiosis, afecta de forma invasiva y potencialmente mortal a individuos inmunodeprimidos (Friedman *et al.,* 2019), aunque es considerada como enfermedad rara, no se descarta para este estudio, dado que recientemente los casos han ido en aumento por causas ya mencionadas.

La **Figura 4.** engloba los linajes fúngicos que contienen patógenos de humanos, ordenados según su división:

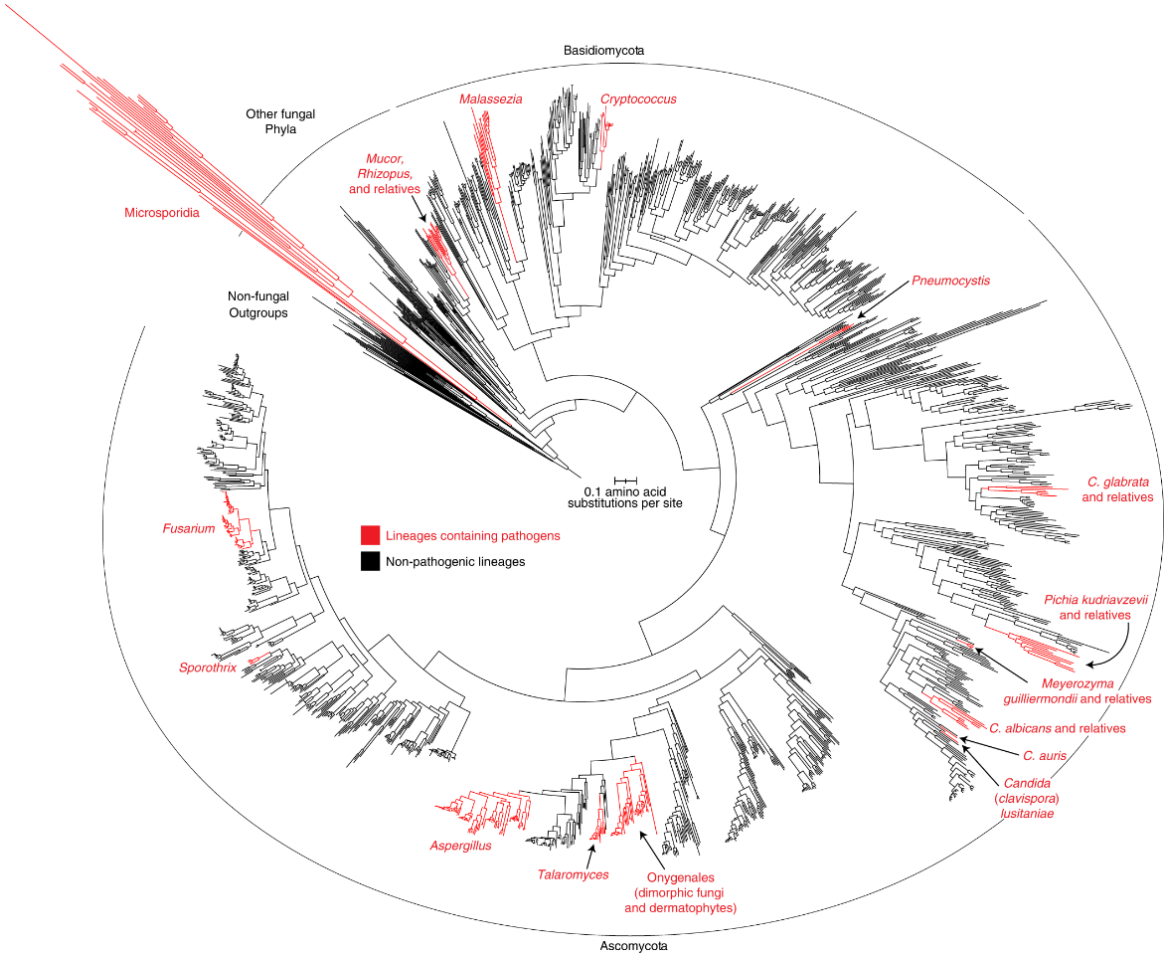

Figura 4. La patogenicidad humana ha evolucionado repetidamente en los hongos.

Los géneros y linajes que albergan patógenos fúngicos principales y emergentes se muestran en rojo y los taxones no patógenos se muestran en negro. Árbol de la vida fúngico basado en un análisis filogenómico de 1644 especies y 290 genes de Li, Y. *et al*., 2021. Solo se incluyen las especies cuyos genomas han sido secuenciados (Rokas, 2022).

2.4 Vía de estrés osmótico

El estrés osmótico conduce a un flujo de agua descontrolado hacia el interior o exterior de la célula: el estrés hiperosmótico causa una contracción, el estrés hipoosmótico causa hinchamiento. La respuesta celular de este tipo de estrés implica la actividad de canales de agua (aquaporinas) y transportadores de electrolitos, y la acumulación de osmolitos (como estrategia de osmorregulación para aumentar la turgencia celular y mantener el metabolismo enzimático) (Mager *et al*., 2000).

Existen factores que desencadenan la respuesta adaptativa como, por ejemplo: el efecto de la humedad (utilizada como porcentaje de contenido de humedad o presión osmótica) en los sustratos, tiene un efecto potencial en la producción de enzimas en estos organismos, así mismo lo es la concentración y disponibilidad de nutrientes en el medio, debido a ello la actividad celular disminuye por lo que el crecimiento vegetativo (conidiación) se ve interrumpido, hasta que el organismo fúngico entra en homeostasis nuevamente (Folch *et al.,* 2004; Martínez *et al.,* 2016; Skoneczny, 2018).

En hongos filamentosos y en otros organismos eucariotas, este estrés se presenta a través de la acción de la vía del glicerol de alta osmolaridad (HOG: high osmolarity glycerol), que lleva a una respuesta adaptativa/osmorregulación en entornos principalmente hiperosmóticos (Furukawa *et al.*, 2005; Skoneczny, 2018).

La vía de HOG es una de las vías de MAP cinasas mejor caracterizadas. Las MAP cinasas son unidades de señalización altamente conservadas en eucariontes, las cuales participan en respuesta a factores ambientales, hormonas, factores de crecimiento o citocinesis. Estas vías controlan el crecimiento celular, la morfogénesis, la proliferación y muchas de las vías de respuesta a estrés (Duran *et al.,* 2010). Sin duda la vía HOG es muy importante tomando en cuenta también que regula más de 80 genes participes en la respuesta al estrés osmótico (Hohmann, S. & Mager, W., 1997) y varios otros relacionados con la conidiación.

Si bien esta vía se encuentra en gran medida conservada, sabemos que hay diferencias para cada especie fúngica dada su naturaleza. Es decir, comparten la respuesta ante un estrés, pero tienden a tener diferencias en las unidades de señalización, como es el caso de *Aspergillus nidulans* (hongo filamentoso) *y Saccharomyces cerevisiae* (levadura) en la vía de estrés osmótico, los cuales presentan proteínas ortólogas*. A. nidulans/ E. nidulans* ha sido un modelo clásico a lo largo de los años para estudios de biología del desarrollo y regulación, esto le ha permitido ser de los hongos mejores caracterizados, por lo que ha sido de elección para la

investigación del genoma fúngico filamentoso (David *et al.,* 2008; Kumar, 2020). Este ascomiceto ha sido seleccionado como el hongo de referencia para este proyecto.

En el estudio por Daisuke *et al.*, 2016, se comparó la dinámica molecular en la vía de estrés osmótico de *A. nidulans y S. cerevisiae/ S. pombe*, **Figura 5**.

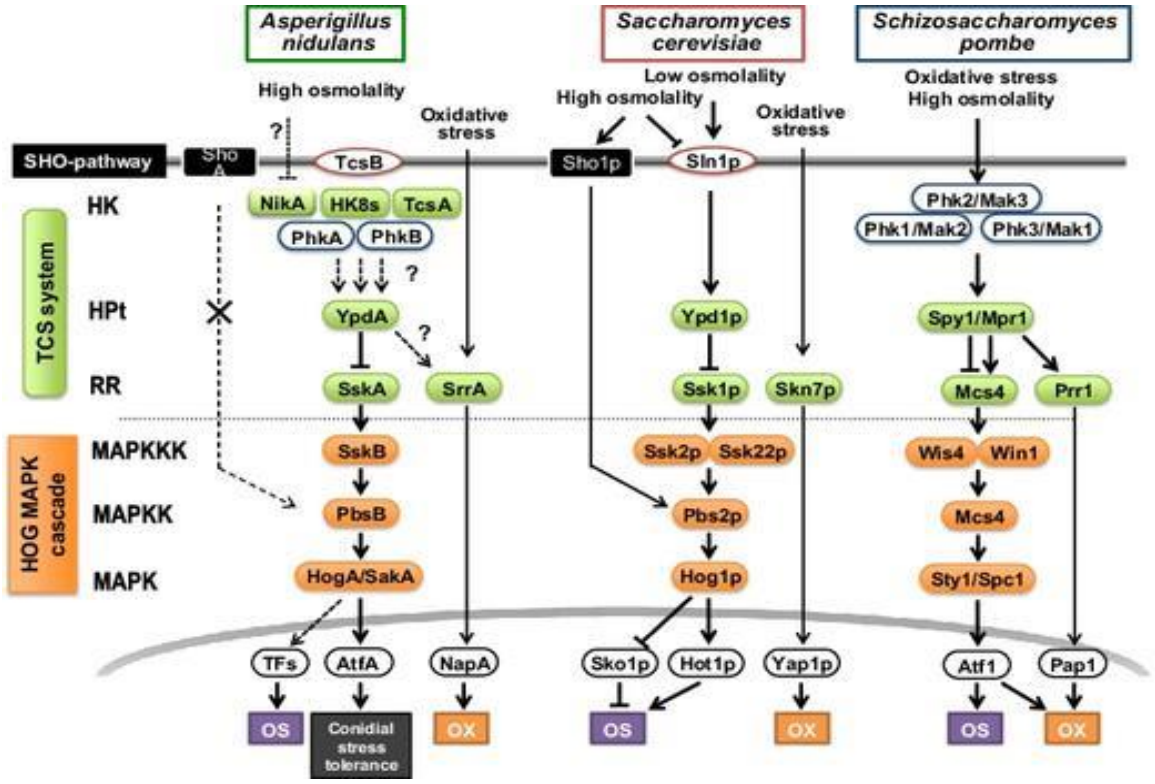

Figura 5. Comparación de componentes de señalización en la vía HOG entre *A. nidulans, Saccharomyces cerevisiae* y *Schizosaccharomyces pombe.*

Se muestran como señales de entrada la osmolaridad y el estrés oxidativo. En *S. cerevisiae*, Sho1p (vía SHO) activa Pbs2p, mientras en la vía homóloga de *A. nidulans* no la activa. Se muestran los HK representativos de *A. nidulans*. La interacción entre HK y YpdA no se ha establecido completamente en *A. nidulans* (flechas punteadas). Las respuestas de salida se representan debajo de la vía de señalización relevante (OS: respuesta al estrés osmótico, OX: respuesta al estrés oxidativo). En *A. nidulans,* la vía de señalización del sistema de señalización de dos componentes (TCS), la cascada HOG MAPK, el factor de transcripción AtfA (TF) están involucradas en la tolerancia al estrés conidial (Daisuke *et al.,* 2016).

La vía HOG de *S. cerevisiae* consta de Ssk2p/Ssk22p MAPKKK, Pbs2p MAPKK y Hog1p MAPK (la cual controla la expresión de los genes biosintéticos de glicerol *GPD1* y *GPP2*), y está regulada por dos ramas aguas arriba diferentes; el sistema TCS y la vía Sho1p, y a diferencia de esta especie, *S. pombe* no presenta la vía Sho1p, y su vía consta de Wis4/Win1 MAPKKK, Mcs4 MAPKK y Sty1/Spc1 MAPK, además, su sistema TCS se compone de tres HK (histidina quinasa): Phk1–Phk3, un HPt (transductor de señal de fosfotransferencia que contiene His): Mpr1 y tres RR (regulador de respuesta): Mcs4, Prr1 y Cek1 (Daisuke *et al*., 2016).

En *S. cerevisiae* el mecanismo de regulación de la vía HOG en condiciones de baja osmolaridad se da de la siguiente manera: el sensor extracelular Sln1 es autofosforilado en un residuo de histidina (H) dentro del dominio Histidina-cinasa, a continuación, el fosfato es transferido a un residuo de ác. aspártico dentro del dominio de respuesta RR, después el fosfato es transferido a Ypd1p en un residuo de histidina y después a Ssk1 en un residuo de ác. aspártico. Se termina la fosforilación e inactiva la vía HOG. Por otro lado, en condiciones de alta osmolaridad Sln1 no es fosforilado por lo que Ssk1 está libre para unirse a Ssk2/Ssk22, después Ssk2 es activado por fosforilación y activa a Pbs2, la cual activa a Hog1, una segunda alternativa es la vía por Sho1p (Daisuke *et al*., 2016; de Nadal, E. & Posas, F., 2022).

Sho1p opera como un interruptor y este es encendido cuando hay un estrés mecánico abrupto, que sucede, por ejemplo: por el desprendimiento de la membrana plasmática de la pared celular, por lo que Sho1p ayuda a conectar señales de la pared celular con la vía HOG (Skoneczny, 2018). Ante una respuesta al estrés osmótico, Sho1p es activado por el osmosensor Msb2 o Hkr1 y después Msb2 activado interactúa con Bem1 citosólico para reclutar Ste20 (proteína quinasa implicada en la regulación de polaridad celular y el ciclo celular) en el complejo, permitiendo la activación de Ste11 (factor de transcripción). La activación de Hkr1 de la vía HOG no requiere Bem1. Ste11 con ayuda de Ste50 proteína adaptadora (modula la especificidad de señalización de MAPK) llega a activar la vía HOG mediante Pbs2, ver **Figura 6** (Brewster, J. L., & Gustin, M. C., 2014; Sharmeen *et al.,* 2019).

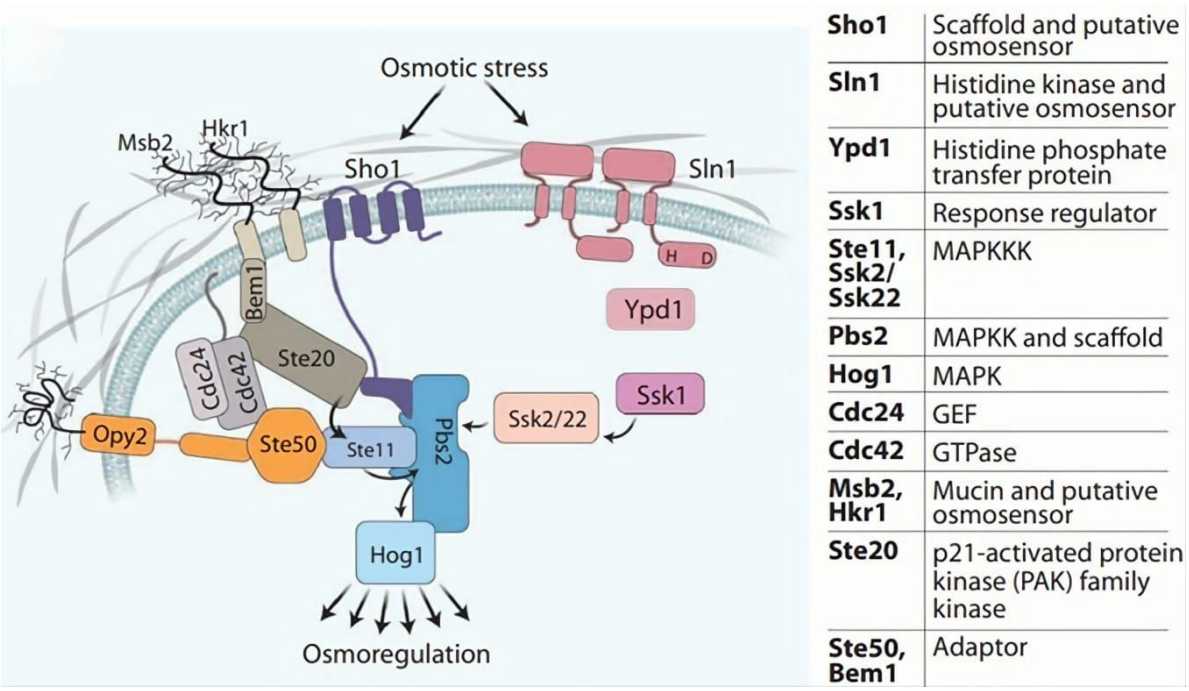

Figura 6. Resumen del modelo actual de la vía HOG en levadura.

El estrés osmótico libera la inhibición dependiente de Sskn1 de Ssk2/22 para activar la vía. La activación de la rama Sho1 requiere que los proteoglicanos de mucina incrustados en la membrana Msb2 o Hkr1 interactúen con Sho1 y Ste20 en un complejo con los componentes MAPK. Opy2 es una glicoproteína transmembrana que sirve como ancla para Ste50. Cdc24 y Cdc42 son la guanosina trifosfatasa citosólica (GTPasa) y el factor de intercambio de nucleótidos de guanina (GEF) que activan Ste20 (Brewster, J. L., & Gustin, M. C., 2014).

En el caso de *A. nidulans*, la vía de señalización por estrés osmótico consta de un sistema TCS también llamado “His-Asp phospho-relay signaling” conformado por TcsB ortólogo del sensor extracelular Sln1, tres HK: NikA (histidina quinasa de tipo Mak2 prescindible en la respuesta al estrés osmótico, la cual transmite señales de estrés inducidas por fungicidas), HK8s y TcsA (quinasa sensora) acopladas a PhkA y PhkB, un HPt: YpdA ortólogo de Ypd1p y un RR: SskA (importante en la regulación de la tolerancia de los conidios al estrés osmótico) y una cascada HOG compuesta por tres proteínas cinasas, SskB, PbsB y HogA/SakA, ortólogos de Ssk2p/Ssk22p, Pbs2p y Hog1p en *S. cerevisiae* (Bahn, 2008). En este caso HogA es ortólogo de Hog1p por lo que controla genes homólogos para la biosíntesis del glicerol: *GfdA, GfdB* y *GppA* (Miskei *et al.,* 2008).

En *A. nidulans*, el mecanismo de regulación de la vía HOG se activa en condiciones de estrés conidial, osmótico u oxidativo. Las proteínas TcsB-YpdA-SskA de *A. nidulans* parecen constituir funcionalmente un sistema de señalización de dos componentes osmosensor similar a las

proteínas de levadura Sln1p-Ypd1p-Ssk1p (Duran *et al.,* 2010). Un punto importante por resaltar es que en *A. nidulans* no se ha encontrado presente la vía de osmorregulación por ShoA que pudiera ser ortóloga de Sho1p de *S. cerevisiae,* por lo que se cree que esté involucrado en otras vías de señalización ya que se ha encontrado que la proteína ModA (importante en la morfogénesis de las hifas) de *A. nidulans* es similar en un 78–94% a Cdc42 (GTPasa) de *S. cerevisiae,* que es un regulador del proceso de la generación de polaridad celular (Virag *et al.,* 2007). Otra proteína es Stec un homólogo de Ste11 (de *S. cerevisiae,* la cual rige el apareamiento, la osmosención y el crecimiento filamentoso), Stec forma un complejo con Ste7 y MpkB y con ayuda de Ste50 media el desarrollo y el metabolismo secundario en *A. nidulans* (Bayram *et al.,* 2012). En la **Figura 7** se muestra el modelo más reciente de la vía de señalización.

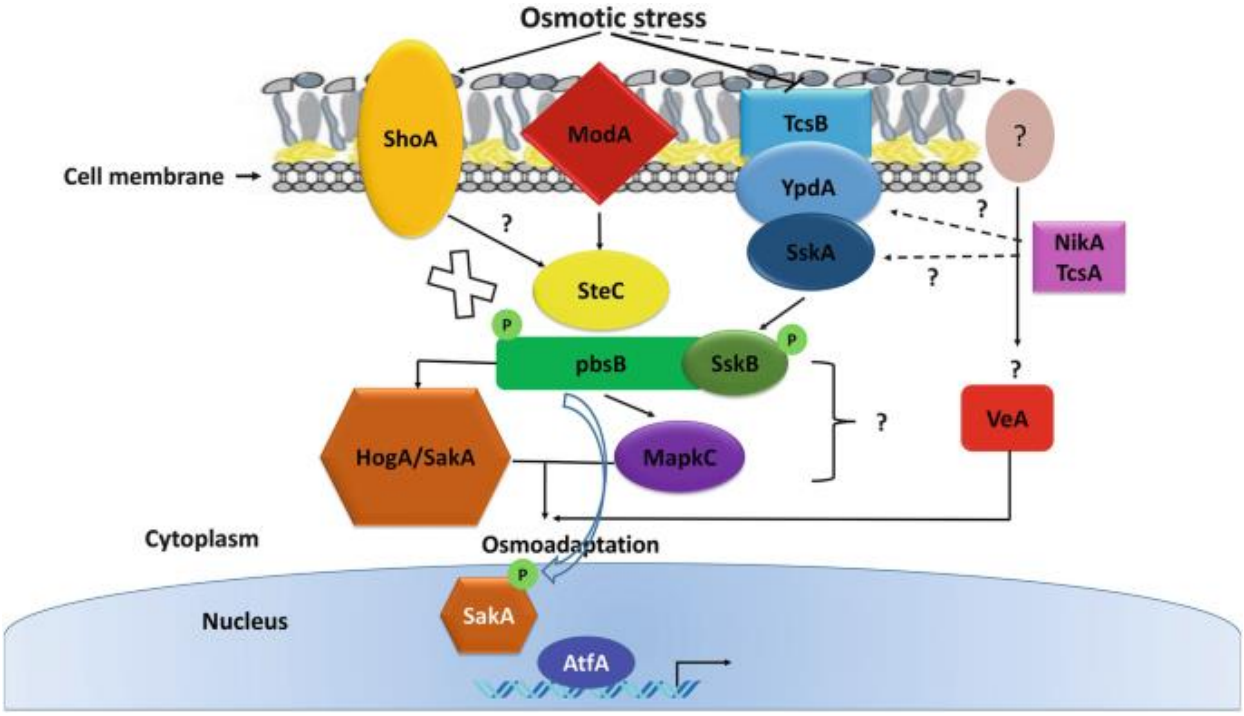

Figura 7. Vía reguladora de HOG del sistema modelo *Aspergillus nidulans* y el papel de la vía de respuesta al estrés osmótico en el desarrollo fúngico y el metabolismo secundario.

La vía HogA (SakA) de *A. nidulans* se activa en la respuesta osmótica y oxidativa por el ortólogo Ssk1 de *A. nidulans*, SskA. La osmorregulación en *A. nidulans* difiere de la de la levadura. Los componentes aguas arriba de esta vía son el homólogo Sln1, TcsB; el ortólogo Ypd1, YpdA; y también NikA (tipo Mak2). *A. nidulans* PbsB MAPKK activa otro ortólogo Hog1, MpkC, que no está presente en la levadura.

Proteínas como ModA, similar a Sho1, y SteC, similar a Ste11, tienen roles en la morfogénesis y el desarrollo sexual. Existen reguladores de transcripción dependientes de SakA en *A. nidulans* como AtfA. Por otro lado, se ha descubierto que VeA, aparte de controlar la morfogénesis sexual / asexual y el metabolismo secundario en hongos filamentosos, participa en la modulación de la conidiación inducida por estrés osmótico, pero poco se sabe sobre el vínculo exacto entre VeA y los componentes de la vía Hog1 de *A. nidulans* (En todas las figuras, las flechas indican regulación positiva, y las líneas con una barra al final indican regulación negativa. Las líneas continuas se utilizan cuando hay evidencia experimental directa y las líneas punteadas cuando el papel aún es hipotético) (Aghcheh, R.K. & Braus, G.H., 2018).

En la actualidad ha incrementado la disponibilidad de secuencias de genomas completos de especies de hongos (en especial los de interés industrial), cada vez es más factible realizar comparaciones bioinformáticas de los reguladores de estrés y por lo tanto también caracterizar de forma óptima las vías de señalización del estrés fúngico. Esto a través de la identificación de ortólogos putativos de las proteínas participes en la vía de señalización, encontrados en los genomas de estos organismos, todo esto gracias a las herramientas bioinformáticas y la proteómica (Muggia *et al.,* 2020).

2.5 Importancia de las herramientas bioinformáticas y la proteómica

La bioinformática es una disciplina científica multidisciplinaria, en la cual se aplican herramientas de computación y análisis a la captura e interpretación de datos biológicos, con el fin de organizar, analizar interpretar, visualizar y almacenar esta información a gran escala. Así es como ayuda a la comunidad científica de todo el mundo a acceder fácilmente a la información, además de crear y/o optimizar herramientas para analizar e interpretar datos de forma masiva (Gauthier *et al.,* 2018; Bayat, 2002).

La proteómica (estudio a gran escala de los proteomas) se enfoca en un estudio de aproximación al funcionamiento y estructura de un conjunto de proteínas que conforman el proteoma. Los objetivos de la proteómica son: El estudio de los cambios globales de la expresión de las proteínas celulares en el tiempo (evolución en la dinámica celular) y la determinación de la identidad y las funciones de todas las proteínas producidas por los organismos (Torrades, 2004).

En sí su investigación se centra en métodos exactos y “rápidos” para identificar y caracterizar proteínas: electroforesis bidimensional en gel y espectrometría de masas, que, con ayuda de

softwares especializados (Andromeda, Mascot, IsobariQ, MaxQuant etc, descritos en Chen *et al*., 2020 y herramientas bioinformáticas, obtienen proteomas de alto rendimiento al manejar de forma masiva y simultanea los datos (Hernández *et al.,* 2019) y obtener por ejemplo de la asignación de espectros las secuencias peptídicas, también estudios *in silico* del reensamblado de los péptidos identificados, la validación y cuantificación a nivel de péptidos y proteínas, anotación y caracterización funcional, así como el almacenamiento de los resultados en una base de datos (Keerthikumar, 2017).

Una de las herramientas bioinformáticas utilizadas es la plataforma BLAST "La Herramienta Básica de Búsqueda de Alineación Local", la cual encuentra regiones de similitud local entre secuencias, que se pueden usar para inferir relaciones funcionales y evolutivas entre secuencias, así como para ayudar a identificar miembros de familias de genes" (NCBI). BLAST se conforma de varios programas de búsqueda por similitud que se han creado para explorar todas las bases de datos de secuencias, estos programas son:

- BLASTP: Compara secuencias de aminoácidos con una secuencia proteica de la base de datos.
- BLASTN: Compara secuencias nucleotídicas con secuencias de ADN de los bancos de datos.
- BLASTX: Compara una secuencia de nucleótidos (traducida a proteínas, con todas sus posibles pautas de lectura) respecto a una proteína de la base de datos.
- TBLASTN: Compara una secuencia proteica con una secuencia nucleotídica de la base de datos traducida a todas sus pautas de lectura.
- TBLASTX: Compara los seis marcos de lectura de una secuencia de nucleótidos respecto las seis pautas de lectura de una secuencia de nucleótidos de la base de datos (McGinnis, S., & Madden, T. L., 2004).

Esta plataforma es de las más utilizadas actualmente, y aunque permite de forma óptima la homología entre proteínas dado el mejor alineamiento posible, no determina la mejor estructura.

Otras plataformas de análisis y de información biológica son: InterPro (para la clasificación de las familias de proteínas), Jalview, Clustal Omega, Expert Protein Analysis System (ExPASy), European Molecular Biology Open Software Suite (EMBOSS), PROSPECT (PROtein Structure Prediction and Evaluation Computer ToolKit), iTOL: el sitio web del Árbol de la vida interactivo para la visualización y manipulación de árboles filogenéticos, desarrollado y mantenido por el

Laboratorio Europeo de Biología Molecular. Protein Data Bank (PDB), Grupow: alineación de secuencias de AND y proteínas, Protein Explorer, etc (Jiang, 2013).

Los lenguajes de programación como Python, Perl, Java, entre otros, se han acoplado a las herramientas bioinformáticas para poder procesar los datos a gran escala y así poder facilitar el análisis. Por ejemplo, el proyecto BioPhython: se conforma de herramientas Python disponibles gratuitamente para biología molecular computacional. Este proporciona el módulo Bio.Blast para manejar la operación NCBI BLAST, puede ejecutar BLAST en una conexión local o a través de una conexión a Internet. Así también cuenta con el módulo Bio.Entrez. (Biophyton).

Existen también sitios web bioinformáticos útiles (disponibles gratuitamente en Internet) (Bayat, 2002; UniProt, NCBI, Pfam):

- Centro Nacional de Información Biotecnológica (**www.ncbi.nlm.nih.gov**): mantiene herramientas y bases de datos bioinformáticas. Por ejemplo: proporciona el navegador Entrez, que es un sistema integrado de recuperación de bases de datos que permite la integración de bases de datos de secuencias de proteínas y ADN.
- Instituto Europeo de Bioinformática (**www.ebi.ac.uk**): centro de investigación y servicios en bioinformática; gestiona bases de datos de datos biológicos.
- Recurso proteico universal “Universal Protein Resource” (**https://www.uniprot.org/**) es un recurso completo para secuencias de proteínas y datos de anotación. Las bases de datos de UniProt son UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef) y UniProt Archive (UniParc) **Figura 8**:

Figura 8. Pestaña principal de UniProt.

Muestra las bases de datos disponibles, así como los proteomas disponibles de las especies; cuenta con alrededor de 22,121 proteomas de referencia, 282,168 proteomas redundantes y 4. 341 proteomas de organismos eucariotas (**https://www.uniprot.org/**).

- Base de datos de familias de proteínas “Pfam” (**http://pfam.xfam.org**) : proporciona un amplio soporte para la clasificación y anotación automatizada de secuencias de proteínas; contiene anotaciones funcionales, referencias bibliográficas y enlaces a bases de datos para cada familia.
- Centro Nacional de Recursos Genómicos (**www.ncgr.org/**): vincula a los científicos con soluciones bioinformáticas mediante colaboraciones, datos y desarrollo de software.
- Genbank (**www.ncbi.nlm.nih.gov/Genbank**): almacena y archiva secuencias de ADN tanto de proyectos de genoma a gran escala como de laboratorios individuales.
- Ensembl (**www.ensembl.org**): base de datos de anotaciones automáticas sobre genomas.
- SWISS-PROT (**www.expasy.org/sprot/**): importante base de datos de proteínas con datos de secuencias de todos los organismos, que tiene un alto nivel de anotación (incluye función, estructura y variaciones) y es mínimamente redundante (pocas copias duplicadas).

Una buena forma de introducirse a la bioinformática es a través del uso de cuadernos digitales interactivos, como el cuaderno digital “Jupyter notebook” que es una interfaz web fácil de configurar y de usar donde sea, el cual puede ser instalado o simplemente utilizarlo de manera online desde su página oficial. Este permite compartir código abierto, análisis de datos, visualizaciones (que pueden ser interactivas), fórmulas matemáticas y otros medios integrados (por ejemplo, videos de YouTube, imágenes y enlaces web), todo en un solo documento que combina componentes interactivos y narrativos. Así permite el análisis de grandes volúmenes de datos, apoya la reproducibilidad y permite visualizaciones interactivas.

La visualización del entorno es prácticamente sencilla, **Figura 9.,** incluye la elección y el control del núcleo de cálculo “Kernel” (el cual se puede parar en cualquier momento): Python, Julia, R., etc., también tiene opciones de guardar la bitácora, insertar o modificar código, ejecutar o detener el proceso para una celda, reiniciar la bitácora, cambiar a texto html, además de incorporar extensiones (Davies *et al*., 2020).

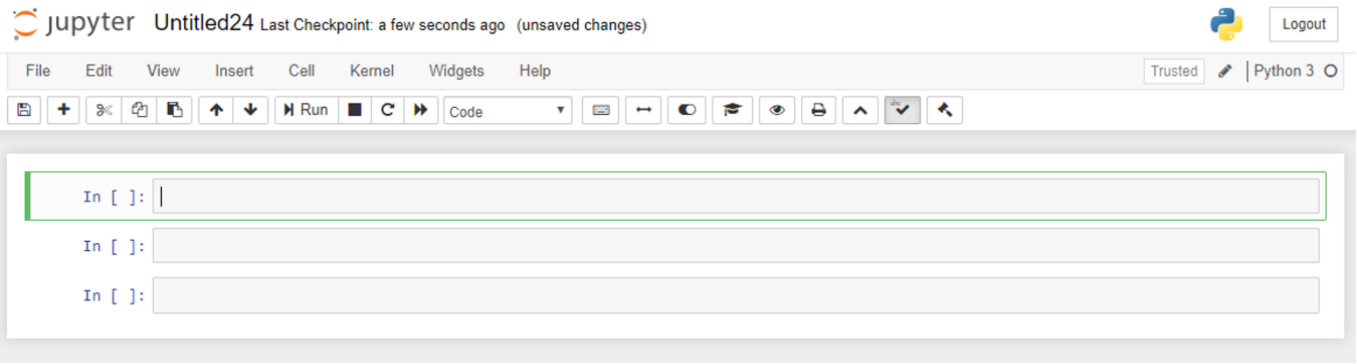

Figura 9. Entorno de Jupyter notebook. La plantilla se conforma del entorno (parte superior) y las celdas (parte inferior) en donde se trabaja y ejecuta el código y permite cambiar a texto para realizar anotaciones (títulos, subtítulos, comentarios, indicaciones etc.) (**https://jupyter.org/**).

3. Objetivo General

Caracterizar mediante un estudio *in silico* la vía de señalización de respuesta a estrés osmótico en hongos filamentosos de la división *Ascomycota*.

3.1 Objetivos específicos

- Realizar una búsqueda exhaustiva de hongos filamentosos de interés industrial y hongos patógenos de plantas, insectos y humanos, en la literatura científica.
- Descarga de proteomas de hongos filamentosos a partir de la base de datos NCBI y UniProtKB.
- Construcción de bases de datos individuales con los proteomas descargados.
- Identificación de proteínas implicadas en la vía de señalización de respuesta a estrés osmótico caracterizadas en especies de *Aspergillus.*
- Identificación de proteínas ortólogas de la vía de señalización mediante el uso del programa Blastp.
- Procesar y representar gráficamente los resultados.

4. Materiales y Métodos

4.1 Selección de hongos filamentosos

Con base en la información mencionada en los antecedentes, para este análisis "*in silico*" se seleccionaron los hongos más relevantes de la división *Ascomycota* (**Tabla 1**).

Tabla 1. Hongos *Ascomycota* de gran interés industrial (2022).

Identificador Taxonómico	Hongo	División	Subdivisión	Clase/Subclase	Orden	Familia	Género
1237896	*Colletotrichum Gloeosporioides*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Hypocreomycetidae*	*Glomerellales*	*Glomerellaceae*	*Colletotrichum*
336722	*Zymoseptoria Tritici*	*Ascomycota*	*Pezizomycotina*	*Dothideomycetes/Dothideomycetidae*	*Mycosphaerellales*	*Mycosphaerellaceae*	*Zymoseptoria*
1268274	*Blumeria Graminis F. Sp. Tritici*	*Ascomycota*	*Pezizomycotina*	*Leotiomycetes*	*Erysiphales*	*Erysiphaceae*	*Blumeria*
431241	*Trichoderma Reesei (Hypocrea Jecorina)*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Hypocreomycetidae*	*Hypocreales*	*Hypocreaceae*	*Trichoderma*
413071	*Trichoderma Virens (Hypocrea Virens)*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Hypocreomycetidae*	*Hypocreales*	*Hypocreaceae*	*Trichoderma*
452589	*Trichoderma Atroviride (Hypocrea Atroviridis)*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Hypocreomycetidae*	*Hypocreales*	*Hypocreaceae*	*Trichoderma*
983964	*Trichoderma Harzianum (Hypocrea Lixii)*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Hypocreomycetidae*	*Hypocreales*	*Hypocreaceae*	*Trichoderma*
1042311	*Trichoderma Asperellum*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Hypocreomycetidae*	*Hypocreales*	*Hypocreaceae*	*Trichoderma*
1290391	*Botryotinia Fuckeliana (Botrytis Cinerea)*	*Ascomycota*	*Pezizomycotina*	*Leotiomycetes*	*Helotiales*	*Sclerotiniaceae*	*Botrytis*
573729	*Myceliophthora Thermophila (Sporotrichum Thermophile)*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Sordariomycetidae*	*Sordariales*	*Chaetomiaceae*	*Thermothelomyces*
857340	*Acremonium Chrysogenum*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Hypocreomycetidae*	*Hypocreales*	*Hypocreales Incertae Sedis*	*Acremonium*
1209962	*Pneumocystis Jirovecii (Human Pneumocystis Pneumonia Agent)*	*Ascomycota*	*Taphrinomycotina*	*Pneumocystidomycetes*	*Pneumocystidales*	*Pneumocystidaceae*	*Pneumocystis*
41688	*Lomentospora Prolificans*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Hypocreomycetidae*	*Microascales*	*Microascaceae*	*Lomentospora*
229533	*Fusarium Graminearum (Gibberella Zeae)*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Hypocreomycetidae*	*Hypocreales*	*Nectriaceae*	*Fusarium*
1397361	*Sporothrix Schenckii*	*Ascomycota*	*Pezizomycotina*	*Sordariomycetes/ Sordariomycetidae*	*Ophiostomatales*	*Ophiostomataceae*	*Sporothrix*
559305	*Trichophyton Rubrum*	*Ascomycota*	*Pezizomycotina*	*Eurotiomycetes/ Eurotiomycetidae*	*Onygenales*	*Arthrodermataceae*	*Trichophyton*

Esta tabla se construyó a partir de los identificadores taxonómicos de la plataforma de UniProt y AspGD. Cada organismo fúngico seleccionado cuenta con un proteoma de referencia, por lo que son aptos para realizar dicho análisis.

Todos los hongos filamentosos se encuentran expuestos a estímulos ambientales y nutricionales que en consecuencia provocan: un estrés fisiológico, el cual altera la homeostasis celular, desencadenando una serie de respuestas que intentan hacer frente a tal desequilibrio. Existen varios tipos de estrés tales como estrés osmótico, oxidativo, lumínico y térmico, y están ligados a procesos fisiológicos de los hongos como el crecimiento radial, conidiación, patogenicidad, y producción de metabolitos secundarios como toxinas, pigmentos y compuestos de interés industrial (Skoneczny,2018).

- Dados los fines del estudio en este trabajo nos enfocamos solo en la vía de respuesta al estrés osmótico.

4.2 Programa Informático e Interfaz

Python v3.6.7: Es un lenguaje de programación de código abierto compatible con Windows, Linux y macOS. Su uso se ha popularizado en la comunidad científica debido a que es un lenguaje poco prolijo y por ende accesible. Se caracteriza por su sintaxis simple en inglés, la cual ayuda a saber qué acción sucederá sin la necesidad de buscar su significado, así como su disponibilidad de módulos y bibliotecas multipropósito.

Python es útil para realizar análisis estadísticos, visualizaciones estadísticas e interactivas, edición de textos, flujos de trabajo, protocolos de investigación y reportes.

- Python fue descargado desde **https://www.python.org/downloads/release/python-367/**, e instalado con permisos de administrador.

Jupyter Notebook: Es una interfaz web dinámica de código abierto estructurada en formato JSON que admite más de 40 lenguajes de programación, incluidos Python, Julia y R, proporciona a los usuarios un entorno apto para crear y ejecutar código, procesar texto con formato, procedimientos matemáticos, imágenes y gráficos, además se puede descargar en formato fasta, txt, HTML, Python, etc. Se utilizó esta interfaz debido a que es flexible y manejable ya que se puede leer el código en todo momento y ejecutar sus partes de forma independiente.

- Jupyter Notebook fue instalado desde la terminal, ejecutando el comando “Python -mpip install jupyter”. **https://jupyter.org/install**.

BLAST v2.8.1: Es la primera versión de producción de BLAST independiente para admitir las nuevas bases de datos BLAST v5 (BLASTDBv5).

- La descarga del programa “ncbi-blast-2.8.1+-win64” se realizó desde Jupyter Notebook, utilizando:

https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.8.1/ncbi-blast-2.8.1+-win64.exe.

Visual Studio Code: Es un editor de código fuente independiente que se ejecuta en Windows, macOS y Linux. Permite la edición, compilación y depuración del código, ya que proporciona soporte integral para la edición, navegación y comprensión del código junto con una depuración ligera, un modelo de extensibilidad enriquecido y una integración ligera con las herramientas existentes. Incluye extensiones para admitir casi cualquier lenguaje de programación (C/C++, Java, Python).

- Visual Studio Code fue descargado desde **https://code.visualstudio.com/Download** e instalado con permisos del administrador, ver instrucciones de instalación para Windows en **https://code.visualstudio.com/docs/setup/windows**.

4.3 Bases de datos

UniProtKB: Es el eje central para la recopilación de información funcional sobre proteínas, con anotaciones precisas, consistentes y ricas. Además de capturar los datos básicos obligatorios para cada entrada de UniProtKB (principalmente, la secuencia de aminoácidos, el nombre o la descripción de la proteína, los datos taxonómicos y la información de citas), se agrega la mayor cantidad de información de anotación posible. Esto incluye ontologías biológicas ampliamente aceptadas, clasificaciones y referencias cruzadas, e indicaciones claras de la calidad de la anotación en forma de atribución de evidencia de datos experimentales y computacionales. Se compone de dos secciones: UniProtKB/Swiss-Prot" (revisado, anotado manualmente) y "UniProtKB/TrEMBL" (no revisado, anotado automáticamente).

- UniProtKB fue consultada en línea desde **https://www.uniprot.org/** y se utilizó tal dirección para la descarga a través de Jupyter Notebook.

La base de datos de UniProtKB fue utilizada para descargar las secuencias de las proteínas involucradas en la vía de señalización de estrés osmótico y los proteomas de los hongos filamentosos.

NCBI protein: La base de datos de proteínas es una colección de secuencias de varias fuentes, incluidas traducciones de regiones de codificación anotadas en GenBank, RefSeq y TPA, así como registros de SwissProt, PIR, PRF y PDB.

- NCBI fue consultada en línea desde **https://www.ncbi.nlm.nih.gov/protein/**

AspGD: Es un recurso genómico en línea que examina la genética y biología molecular de *Aspergillius* y tiene un enfoque genómico comparativo para refinar y mejorar iterativamente las anotaciones de genes estructurales en múltiples especies *Aspergillius.*

- AspGD fue consultada en línea desde **http://www.aspergillusgenome.org/**

AspGD fue utilizada para cotejar que los identificadores correspondieran a cada uno de los proteomas buscados.

5. Resultados y Discusión

5.1 Identificación de proteínas implicadas en la vía de señalización de respuesta a estrés osmótico caracterizadas en especies de *Aspergillus*

A partir de la base de datos de UniProt, NCBI protein y AspGD, se identificaron las 21 proteínas que participan en la vía de señalización, así mismo a través de **https://www.ebi.ac.uk/interpro/** se logró visualizar y extraer su arquitectura. Su importancia radica en que a partir de ello se puede realizar un correcto alineamiento entre las secuencias (*E. nidulans* (hongo modelo de referencia) vs hongos filamentosos) en el que se busca encontrar su ortólogo u ortólogos, considerando: A) que pertenecen a la misma división (*Ascomycota*), B) que comparten características generales, C) que deben tener los mismos dominios funcionales, ya que esto indica la funcionalidad de la proteína y nos da una idea del grado de conservación de la proteína y D) la longitud de la proteína debe ser altamente similar o igual a la otra, ya que ello nos habla de la composición. Estos aspectos son fundamentales para la obtención de una alta puntuación de alineamiento en el estudio *in silico* y por ende en la identificación de proteínas potencialmente ortólogas en diferentes hongos filamentosos pertenecientes a la división *Ascomycota.*

Las proteínas identificadas en *Aspergillus* pertenecen al sistema de dos componentes: HK, HPt y RR y la cascada HOG MAPK: MAPKKK, MAPKK y MAPK, y GTPasas, todas ellas implicadas en la respuesta a estrés osmótico, aunque algunas de ellas también participan en otras vías de señalización que responden a distintos estímulos. Por ejemplo: varias de ellas se han visto involucradas en la respuesta a estrés oxidativo, la producción de conidios (desarrollo sexual), la inhibición por fungicidas, la resistencia al estrés de la pared celular, en el proceso de la generación de polaridad celular, siendo algunas de ellas: AtfA, napA, steC, steD, ste20, modA.

Aun cuando las vías de señalización están relativamente conservadas, no son idénticas en el género *Aspergillus* (de Vries *et al*., 2017). Un ejemplo son las diferencias entre el papel de *A. nidulans* SakA y SakA del patógeno humano *A. fumigatus*, en respuesta a las tensiones. En *A. fumigatus* SakA es crucial para la respuesta al estrés oxidativo y la producción de metabolitos secundarios, mientras que en *A. nidulans* SakA apenas tiene impacto en la osmorregulación (Aghcheh, R.K. & Braus, G.H., 2018).

En la **Figura 10**. podemos visualizar las proteínas que caracterizan a la vía de estrés osmótico en especies de *Aspergillus* y también algunas que participan en esta y otras vías.

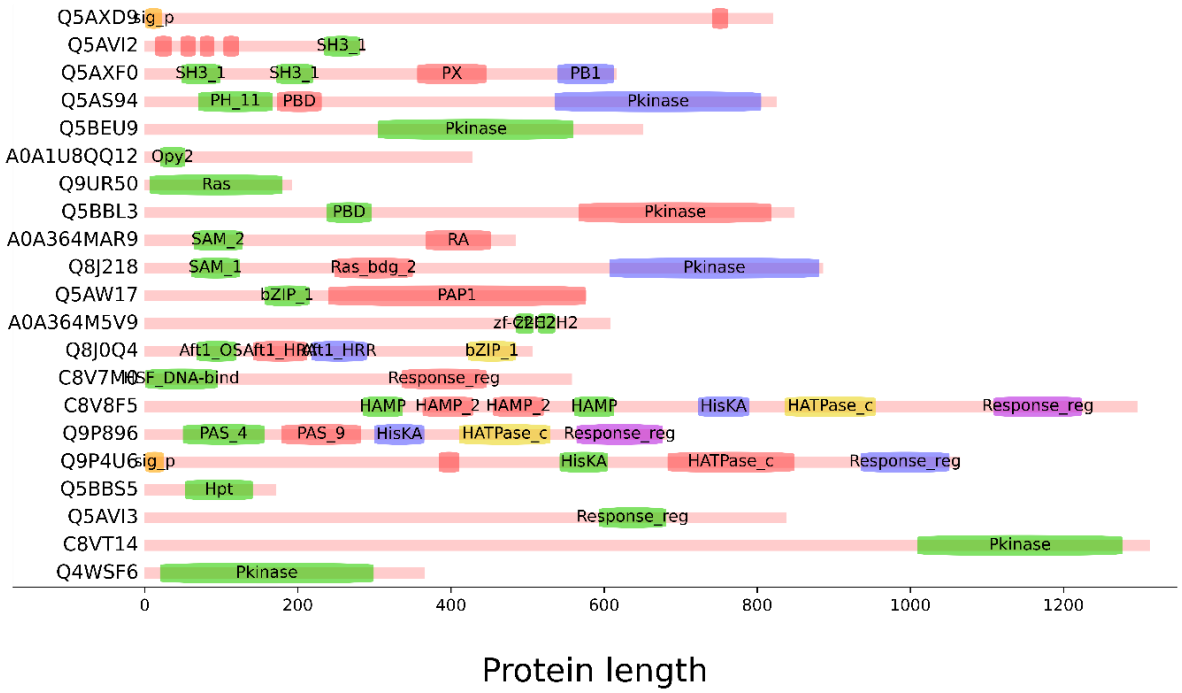

Figura 10. Proteínas implicadas en la vía de señalización de respuesta a estrés osmótico caracterizadas en especies *Aspergillus.*

Datos de UniProt y Pfam/InterPro.

- **Q4WSF6.** Hog1_ASPFU (*A. fumigatus*): Proteína quinasa activada por mitógenos involucrada en una vía de transducción de señales que se activa por cambios en la osmolaridad del entorno extracelular. Controla la regulación osmótica de la transcripción de genes diana. Sinónimos: **hogA**, osm1**, sakA.** Gen: ***hog1*.** aa: **366.**
- **C8VT14.** C8VT14_EMENI (*A. nidulans*): Proteína MAP quinasa quinasa quinasa. Dominio: proteína quinasa. Gen: ***sskB***, *ANIA_10153*. aa: **1313**.
- **Q5AVI3.** Q5AVI3_EMENI (*A. nidulans*): Regulador de la respuesta citoplasmática que interactúa con HK fosforiladas y cataliza la transferencia de un grupo fosforilo y cataliza la autodesfosforilación. Dominio: regulador de respuesta. Gen: ***sskA****, ANIA_07697*. aa: **838.**
- **Q5BBS5.** Q5BBS5_EMENI (*A. nidulans*): Proteína de fosfotransferencia que contiene histidina. Dominio: HPt que tiene un doble propósito como receptor de fósforo y donante de fósforo para transportar grupos fosforilo entre dos o más dominios reguladores de respuesta. Gen**: *ypdA*,** *ANIA_02005*. aa: **172.**
- **Q9P4U6.** TCSB_EMENI (*A. nidulans*): Proteína B del sistema de dos componentes que se activa por cambios en el entorno extracelular. Dominios: Histidina quinasa, HATPasa o ATPasa similar a la histidina quinasa, regulador de respuesta. Gen: ***tcsB***, *AN1800*. aa: **1065.**
- **Q9P896.** TCSA_EMENI (*A. nidulans*): Proteína A del sistema de dos componentes que se activa por cambios en el entorno extracelular. Dominios: PAS, Histidina quinasa y regulador de respuesta. Gen: ***tcsA***, *AN5296*. aa: **682.**
- **C8V8F5.** C8V8F5_EMENI (*A. nidulans*): Proteína histidina quinasa. Dominios: HAMP se sugiere que posee la función de regular la fosforilación, Histidina quinasa, HATPasa y regulador de respuesta. Gen: ***nikA***, *ANIA_04479.* aa: **1297.**
- **C8V7M0.** C8V7M0_EMENI (*A. nidulans*): Regulador de respuesta al estrés.
  Dominios: Factor de choque térmico (HSF) activador transcripcional de los genes de choque térmico y regulador de respuesta. Gen: ***srrA***, *ANIA_03688*. aa: **558.**
- **Q8J0Q4.** Q8J0Q4_EMEND (*A. nidulans*): Factor de transcripción AtfA que interactúa con MAPK SakA para regular las respuestas generales al estrés, el desarrollo y las funciones de las esporas. Dominios: dominio de estrés osmótico, HRA, HRR y BZIP. Gen: ***atfA*.** aa: **507.**
- **A0A364M5V9.** A0A364M5V9_ASPFL (*A. flavus*): Factor de transcripción C2H2 (**Seb1**). Dominios: tipo C2H2 de dedo de zinc, se sugiere que tiene interacción con nucleótidos. Gen: ***msnA***. aa: **608.**
- **Q5AW17**. AP1_EMENI (*A. nidulans*): Activador de la transcripción implicado en la respuesta al estrés oxidativo, específicamente durante el crecimiento de las hifas. Dominio: bZIP. Factor de transcripción: PAP1 que regula la transcripción de genes antioxidantes en respuesta a H2O2. Gen: ***napA*.** aa: **577.**
- **Q8J218.** Q8J218_EMEND (*A. nidulans*): MAPKK quinasa. Dominios: proteína quinasa, motivo alfa estéril (SAM), Dominio de unión a Ras de Byr2: responde a la señalización de feromonas y controla el apareamiento a través de una vía MAPK. Gen: ***steC*.** aa: **886**.
- **A0A364MAR9.** A0A364MAR9_ASPFL (*Aspergillus flavus):* MAPKKK cascada proteína quinasa regulador Ste50. Dominios: Dominio asociado a Ras (RA_dom), las proteínas Ras son GTPasas transductoras de señales que alternan entre formas inactivas unidas a GDP y formas activas unidas a GTP y dominio de motivo alfa estéril (SAM). Gen: ***steD*.** aa: **485**.
- **Q5BBL3.** STE20_EMENI (*A. nidulans*): Componente MAP4K de la vía MAPK requerido para la respuesta de feromonas de apareamiento y la regulación de la polaridad celular y el ciclo celular. Dominios: Dominio CRIB, también llamado dominio de unión de p21 (PBD), que se ha demostrado que se une específicamente a la forma unida a GTP de Cdc42 o Rac, con preferencia por Cdc42. Dominio: proteína quinasa. Gen: ***ste20*.** aa: **848.**
- **Q9UR50.** Q9UR50_EMEND (*A. nidulans*): GTPasa homóloga de la proteína 42 de control de la división celular. Pertenece a la subfamilia Cdc42. Dominio: Ras. Gen: ***modA*.** aa: **192**.

- **A0A1U8QQ12**. A0A1U8QQ12_EMENI (*A. nidulans*): Proteína que contiene el dominio **Opy2** la cual, actúa como un ancla de membrana en la vía de señalización HOG. Gen: ***AN2522.2.*** aa: **428**.
- **Q5BEU9.** Q5BEU9_EMENI (*A. nidulans)*: MAPKK Dominio: proteína quinasa. Gen: ***pbsA.*** aa: **651**.
- **Q5AS94**. Q5AS94_EMENI (*A. nidulans*): Proteína quinasa. Dominios: proteína quinasa, CRIB o PBD (dominio de unión PAK o quinasa activada por p21) se une a pequeñas GTPasas similares a Cdc42p y/o Rho. Subfamilia STE20. Gen: ***ANIA_08836***. aa: **825**.
- **Q5AXF0**. Q5AXF0_EMENI (*A. nidulans*): Activador de proteína quinasa Bem1(**BemA**). Dominios: SH3 media en el ensamblaje de complejos proteicos específicos mediante la unión a péptidos ricos en prolina. PX. Gen: ***ANIA_07030***. aa: **616.**
- **Q5AVI2**. SHO1_EMENI (*A. nidulans*): Osmosensor de membrana plasmática que activa la vía de señalización MAPK del glicerol de alta osmolaridad (HOG) en respuesta a la alta osmolaridad. Las altas concentraciones de sal conducen a la localización en la membrana de MAPKK Pbs2, que luego es activada por MAPKKK Ste11 y, a su vez, activa MAPK Hog1. Pbs2 se localiza en la membrana a través de la interacción de su motivo PxxP con el dominio SH3 de Sho1. Dominio: SH3. Gen: ***sho1.*** aa: **287.**
- **Q5AXD9**. Q5AXD9_EMENI (*A. nidulans*): Proteína de señalización de la familia de mucina **Msb2**. En la activación de la cascada Hog se requiere la proteína de andamiaje Bem1 y el citoesqueleto de actina, lo que sugiere que Msb2 funciona como un osmosensor para integrar las señales de las condiciones osmóticas externas con las de las condiciones internas del citoesqueleto. Gen: ***ANIA_07041***. aa: **821.**

5.2 Construcción de una base de datos BLAST

Se realizó una base de datos BLAST con las proteínas implicadas en la vía de señalización de respuesta a estrés osmótico caracterizadas en especies de *Aspergillus* (**Figura 11**):

Figura 11. Proteínas implicadas en la vía de señalización de respuesta a estrés osmótico en especies de *Aspergillus.*

Su localización se realizó a través de su identificador en UniProt, NCBI y AspGD como plataformas auxiliares.

Se creó el archivo “stress.fasta” a partir de las secuencias de las proteínas descargadas en formato FASTA desde UniProtKB.

Nota: Los archivos FASTA se modificaron y se concatenaron en un solo archivo, en los encabezados solo quedo el identificador, esto con la finalidad de mejorar el manejo de datos.

Después se ejecutó la celda de código para la creación de la base de datos “DB2/stress” (**Figura 12)**.

```
1 subprocess.call('makeblastdb -in stress.fasta -dbtype prot -parse_seqids -out DB2/stress', shell = True)
```

```
0
```

Figura 12. Construcción de una base de datos BLAST a partir de secuencias de referencia en formato fasta.

El módulo “subprocess” llama al programa BLAST, a continuación, se construye la base de datos de las proteínas de interés mediante “makeblastdb”, a partir del multifasta. “in”: el multifasta (“final2_Larissa.fasta”) con las secuencias que forman la base de datos, “dbtype prot”: tipo de datos:

proteínas, “parse_seqids”: Identificadores de secuencia delimitados por barras de análisis (p. ej., gi|129295) en la entrada FASTA. “out”: el nombre que se le puso a la base de datos, en este caso es “final2_Larissa”” y se sitúa en la carpeta DB2.

5.3 Descarga de proteomas a partir de UniProtKB

Para la obtención de los archivos FASTA de los hongos filamentosos seleccionados, se construyó una lista (nombrada “io”) de sus identificadores taxonómicos y nombres (ver **Tabla1,** página 24), incluyendo el hongo modelo “*E. nidulans*” con ID: 227321, los cuales fueron tomados de la página de UniProtKB. La descarga de proteomas se realizó de forma automática sometiendo la variable “io” a un proceso iterativo.

Se importó la librería “requests” con la finalidad de realizar una solicitud http (servidor web) para extraer los proteomas.

En el proceso de extracción de proteomas, se construyó un bucle “for loop”, en el cual la variable “i” tomó y guardó temporalmente uno por uno los elementos de la lista “io”, se crearon las variables “x“ y “d” para extraer los datos en forma de tabla: id, nombre, organismo, cantidad de proteínas y el Detector de proteoma completo (CPD), clasificando cada proteoma en alguna categoría con el fin de conocer la integridad y calidad del proteoma., en la variable “d” se creó un Pandas DataFrame, el cual es una estructura de datos tabulares bidimensionales de tamaño mutable, con ejes como columnas y filas, es decir que los datos se alinean de forma tabular en columnas y filas (pandas v1.1.5, 2022). Para la creación del DataFrame es importante haber importado antes las librerías (Ver **Anexo 1**). Los datos se guardaron en la variable “tab”, ver **Figura 13.**

```
tab = [ ]
for i in io:
    print(i)
    x = requests.get('https://www.uniprot.org/proteomes/?query='+i+'&format=tab&columns=id,name,organism-id,proteincount,cpd').content.decode()
    d = DataFrame([i.split('\t') for i in x.split('\n')][1:-1])
    tab.append(d)
```

Figura 13. Extracción de proteomas a través de UniProtKB.

En la primera línea creó una lista vacía “tab=[ ]”, en la línea 2, el bucle for en el que “i” va guardando temporalmente todos los datos de la lista “io” (identificador, organismo), en la línea 3, va mostrando cada

id del cual se ha extraído la información solicitada, en la línea 4, se utilizó “requests.get” para realizar la solicitud a la página web de UniProt y tomar los datos guardados en la variable “i” y extraerlos de la web, en la línea 5 se construyó el DataFrame a partir de un bucle for y se eliminaron los espacios, por último se anexó el DataFrame a “tab” mediante “.append”.

Los datos obtenidos se muestran en la **Tabla 2**., cabe mencionar que CPD usa información de linaje taxonómico para identificar el grupo de proteomas taxonómicamente más cercano a él. La mayoría muestran un valor Outlier (low value) o Valor atípico (valor bajo), es decir que estos proteomas tienen un recuento por debajo del promedio del grupo y no son tan cercanos a la mediana (estándar) del grupo. Se utilizó la **Tabla 2.** para la descarga de los proteomas en formato fasta a través de un bucle for y el módulo “urllib.request” permitió abrir la URL, evitando redireccionamientos, cookies, etc. Ver **Figura 14.**

Tabla 2. Visualización tabular de la extracción y dimensionalidad de datos proteómicos requeridos a través de UniProtKB.

	Proteome ID	Organism	Organism ID	Protein count	CPD
0	UP000015530	Colletotrichum gloeosporioides (strain Cg-14) ...	1237896	16388	Close to standard (high value)
1	UP000008062	Zymoseptoria tritici (strain CBS 115943 / IPO3...	336722	10972	Outlier (low value)
2	UP000053110	Blumeria graminis f. sp. tritici 96224 (Strain...	1268274	4024	Outlier (low value)
3	UP000008984	Hypocrea jecorina (strain QM6a) (Trichoderma r...	431241	9114	Outlier (low value)
4	UP000007115	Hypocrea virens (strain Gv29-8 / FGSC 10586) (...	413071	12389	Standard
5	UP000005426	Hypocrea atroviridis (strain ATCC 20476 / IMI ...	452589	11815	Standard
6	UP000241690	Trichoderma harzianum CBS 226.95 (Strain: CBS ...	983964	14049	Outlier (low value)
7	UP000240493	Trichoderma asperellum CBS 433.97 (Strain: CBS...	1042311	12547	Outlier (low value)
8	UP000012045	Botryotinia fuckeliana (strain BcDW1) (Noble r...	1290391	11022	Outlier (low value)
9	UP000007322	Myceliophthora thermophila (strain ATCC 42464 ...	573729	9079	Close to standard (low value)
10	UP000029964	Acremonium chrysogenum (strain ATCC 11550 / CB...	857340	8899	Close to standard (low value)
11	UP000010422	Pneumocystis jirovecii (strain SE8) (Human pne...	1209962	3419	Unknown
12	UP000233524	Lomentospora prolificans (Strain: JHH-5317)	41688	8530	Close to standard (low value)
13	UP000070720	Gibberella zeae (strain ATCC MYA-4620 / CBS 12...	229533	14162	Standard
15	UP000033710	Sporothrix schenckii 1099-18 (Strain: 1099-18)	1397361	10292	Close to standard (low value)
16	UP000008864	Trichophyton rubrum (strain ATCC MYA-4607 / CB...	559305	10006	Outlier (low value)
17	UP000000560	Emericella nidulans (strain FGSC A4 / ATCC 381...	227321	10560	Standard

La tabla se encuentra organizada por ID del proteoma, organismo, ID del organismo, proteínas contabilizadas y CPD: Complete Proteome Detector (Detector del Proteoma Completo) el cual nos habla de la integridad y calidad de cada proteoma comparándolo directamente con los de un grupo de al menos tres especies estrechamente relacionadas taxonómicamente (UniProt).

```
for i in up_tax:
  print(io[up_tax[i]], '=', i)
urllib.request.urlretrieve('https://www.uniprot.org/uniprot/?query=proteome:'+i+'&format=fasta','fastas2/'+up_tax[i]+'.fasta')
```

Figura 14. Descarga de proteomas a través de UniProtKB.

Los datos de la lista up_tax que contiene “Proteome ID” y “Organism ID” de la Tabla 2., se fueron guardando temporalmente en la variable “i”, el módulo urllib.request.urlretrieve llamó a la página web de uniprot.org y se fueron descargando cada uno de los proteomas en formato FASTA, los cuales se nombraron de acuerdo a la lista “up_tax” y se guardaron en la carpeta “fastas2”.

5.4 Identificación de proteínas ortólogas de la vía de señalización de respuesta a estrés osmótico mediante Blastp

Los ortólogos: son genes en diferentes especies que evolucionaron a partir de un gen ancestral común por especiación y, en general, los ortólogos conservan la misma función durante el curso de la evolución (Gennarelli, M., & Cattaneo, A., 2010).

El programa Blastp es capaz de comparar una secuencia problema (query) contra una gran cantidad de secuencias que se encuentren en una base de datos (subject) y permite encontrar posibles ortólogos. En este análisis se realizaron varios filtros con el fin de identificar de manera óptima cada proteína ortóloga para cada proteoma.

En el primer filtro (**Figura 15**) se obtuvieron las cinco secuencias de proteínas hits o con mayor similitud en alineamiento con cada proteína implicada en la vía de señalización de estrés osmótico, a través del comando: -max target segs 5 -max hsps 1-. Se ordenó el archivo con base en los especificadores de formato “qacc sacc qlen slen length qstart qend sstart send score bitscore evalue pident nident mismatch positive gaps gapopen” para usar posteriormente en el DataFrame. El resultado se guardó en “Tables2” en formato .txt.

```
for input_file in sec:
  print(input_file)
  subprocess.call('blastp -db DB2/stress -query fastas2/'+input_file+\
                  ' -evalue 1E-6 -outfmt "6 qacc sacc qlen slen length qstart qend sstart send score bitscore evalue pident nident mismatch positive gaps gapopen" -max_target_seqs 5 -max_hsps 1 -out Tables2/'+input_file.split(".fasta")[0]+".txt", shell = True)
```

Figura 15. Código para la identificación de proteínas con mayor similitud.

Este filtro permite obtener los 5 mejores hits para cada posible proteína ortóloga en cada proteoma de los hongos filamentosos de mayor interés, descartando aquellas secuencias con nula o menor similitud. qacc= identificador del query, sacc= identificador del subject, qlen= longitud de la secuencia de consulta, slen= longitud de la secuencia de sujetos , length= longitud de alineación, qstart= inicio de alineación en la consulta, qend= fin de la alineación en la consulta, sstart= inicio de alineación en el sujeto, send= fin de la alineación en el asunto, score= puntuación, pident= porcentaje de coincidencias idénticas, nident= número de coincidencias idénticas, mismatch= número de desajustes, positive= número de partidos con puntuación positiva, gaps= número total de brechas, gapopen= número de espacios abiertos. (Scholz, 2022, BLAST).

Se guardaron los datos de la variable “Tables2” en la variable “es” y se prosiguió a la creación de un Pandas DataFrame, esto permitió analizar los datos a través de su agrupación y organización.

Se construyó un bucle “for loop”, en el cual la variable “a” tomó y guardó temporalmente uno por uno los datos de la variable “es”, el comando “pd.read_csv” leyó los datos de “Tables2” (dado el formato .txt) y se asignaron las columnas, incluyendo “tax” que representa el identificador taxonómico para cada hongo filamentoso, para una mejor visualización. El DataFrame se guardó en una variable “data”, **Figura 16.**

```
data=[]
for a in es:
    ab=pd.read_csv('Tables2/'+a,sep='\t',names=columna)
    ab["tax"]=a.split(".txt")[0]
    data.append(ab)
```

Figura 16. Comando para extraer y organizar información del archivo “Tables2”: proteínas con mayor similitud.

sep= separador, el valor predeterminado es ',' como en CSV, CSV= valores separados por coma “Comma-Separated Values”, \t= carácter de tabulación, columna="qacc", "sacc", "qlen", "slen", "length", "qstart", "qend", "sstart", "send", "score", "bitscore", "evalue", "pident", "nident", "mismatch", "positive", "gaps", "gapopen", split= divide una cadena en una lista, append= agrega elementos al final de la lista.

A continuación, el DataFrame “data” se concatenó (se unieron los datos) y se guardó en la variable “blastp”, después se descartaron aquellos valores de % de identidad menores a 40% (vistos en la columna “pident”), ya que representan valores muy bajos de alineación de

secuencias entre “qacc” y “sacc “. En este primer ejercicio: filtrado de datos del DataFrame, se observó (**Tabla 3**) que más de una proteína del “qacc” había hecho match con una proteína del “sacc”, ya que al realizar la comparación entre secuencias hubo una cierta similaridad (reflejada en él % de identidad) entre fragmentos de aminoácidos perteneciente a regiones proteicas como por ejemplo; dominios y residuos, como era de esperarse, dado que los hongos filamentosos presentan estas regiones conservadas en las cascadas de MAPK (MAPKKK, MAPKK, MAPK) y otras vías de señalización (Rispail *et al.,* 2009).

Tabla 3. Blastp.

```
1 blastp
```

	qacc	sacc	qlen	slen	length	qstart	qend	sstart	send	score	bitscore	evalue	pident	nident	mismatch	positive	gaps	gapopen
9	A0A2T3Z4F9	Q5BBL3	715	848	73	296	368	636	707	147	61.2	1.560000e-12	41.096	30	42	44	1	1
16	A0A2T3ZGR0	C8VNJ9	338	580	50	256	304	462	511	120	50.8	6.240000e-10	46.000	23	26	28	1	1
34	A0A2T3ZB84	Q5BBL3	627	848	282	11	288	567	845	559	219.0	9.340000e-65	40.426	114	161	168	7	4
42	A0A2T3YZ29	C8V8F5	1377	1297	244	792	1031	708	949	476	187.0	5.680000e-51	44.262	108	130	145	6	3
46	A0A2T3Z567	C8V7M0	740	558	84	166	249	1	82	193	79.0	3.690000e-18	41.667	35	47	54	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
570	A0A2T4A8X7	C8VNJ9	843	580	65	24	86	461	525	145	60.5	3.460000e-12	44.615	29	34	40	2	2
579	A0A2T4ARX7	C8VNJ9	721	580	60	500	558	463	522	136	57.0	2.890000e-11	40.000	24	35	34	1	1
587	A0A2T3ZXY7	C8VNJ9	123	580	51	18	67	464	514	89	38.9	3.150000e-07	43.137	22	28	29	1	1
593	A0A2T4A3Q4	G5EB26	908	886	923	3	905	1	882	2207	854.0	0.000000e+00	50.921	470	392	610	61	20
607	A0A2T4AF13	Q5AS94	267	825	79	84	158	620	693	99	42.7	1.700000e-07	40.506	32	38	42	9	3

898 rows × 19 columns

Se utilizó la instrucción “blastp = blastp[blastp.pident >= 40]” para excluir aquellos alineamientos de menor significancia (menor que 40%) entre las secuencias de cada proteína de cada proteoma y las secuencias (referencia) de la base de datos creada. En las celdas punteadas se observa que para la proteína: Factor de transcripción C2H2 (Seb1) de *A. nidulans*, hay 3 proteínas (A0A2T4A8X7, A0A2T4ARX7 y A0A2T3ZXY7) del hongo *Trichoderma harzianum* que son similares a Seb1 ya que presentan el dominio C2H2, sin embargo, la proteína A0A2T4A8X7 tiene un % de identidad= 44.615, el cual es mayor en comparación a las otras 2 proteínas.

Nota: A partir de ello se consideró realizar otros filtros para descartar falsos positivos además considerar aquellas proteínas con la misma arquitectura estructural. Se generó el archivo fasta:” blastp_sequences_Larissa.fasta” donde se guardaron todos los datos del Blastp, como un resguardo de información.

El segundo filtro se realizó al generar un DataFrame por números taxonómicos del blastp, correspondiente a cada hongo filamentoso y al de referencia “*E. nidulans*”, tuvo la finalidad de asegurar el match entre las secuencias de proteínas de cada hongo filamentoso y al menos una proteína de las 21 proteínas de referencia (*E. nidulans*).

Se descartaron aquellas proteínas que no cumplieron con un % de identidad mayor o igual a 50 (umbral de >=50) y a partir de estos datos se generó un nuevo DataFrame del cual se tomó la columna sacc (de la base de datos de referencia) para buscar los identificadores y crear mini DataFrame (mdf) de cada uno, con el fin de concatenarlos y reordenar el índice de la tabla (genera un orden) (**Figura** 17).

```
c=[]
for i in blastp.tax.unique().tolist():
    pr1=blastp[blastp.tax==i]
    pr2=pr1[pr1.pident >=50].reset_index(drop=True)
    if len(pr2) == 0:
        pass
    else:
        lista1=pr2.sacc.unique().tolist()
        data1=[]
        for i in lista1:
            mdf=pr2[pr2.sacc==i]
            data1.append(mdf)
        con=pd.concat(data1).reset_index(drop=True)
        c.append(con)
```

Figura 17. Dataframe para concatenar datos filtrados (>=50) e identificadores de la base de datos.

Primero se realizó la iteración de una lista de no. taxonómicos únicos y se buscó cada uno de no. taxonómicos en la variable blastp (contiene todos los no.tax) y se generó 1 dataframe, a partir de dataframe anterior (pr1), se filtró por % de identidad (por un umbral de >=50) y se generó de nuevo un dataframe (pr2). A partir del pr2 se creó una lista única de la columna sacc (de la base de datos de referencia), se buscó cada uno de sacc (identificadores de la base de datos) en la variable pr2 y se generó un mini dataframe, cada uno de los mini dataframes es guardado en la variable data1(data1 fue definido anteriormente como una lista vacía). Todos los mini dataframe guardados en data1 fueron concatenados y el índice fue reordenado, al finalizar se construyó un nuevo dataframe(con), todo se guardó en la variable “c”.

Este paso se guardó en la variable “table1” a través de ejecutar: table1=pd.concat(c).reset_index(drop=True) (**Tabla 4**).

Tabla 4. DataFrame “table 1; Blastp 1 filtro >=50”. Vista previa; su dimensión es de 312 filas por 19 columnas.

```
In [36]: table1
         #blastp 1 filtro >=50
```

	qacc	sacc	qlen	slen	length	qstart	qend	sstart	send	score	bitscore	evalue	pident	nident	mismatch	positive	gaps	gapopen	tax
0	A0A2T3Z8Y6	Q5BBS5	148	172	135	11	144	24	158	355	141.0	1.690000e-46	55.556	75	59	94	1	1	1042311
1	A0A2T3ZK40	G5EAY4	204	192	199	6	204	3	192	665	260.0	2.060000e-92	60.804	121	69	149	9	2	1042311
2	A0A2T3YVF3	G5EAY4	194	192	194	1	194	1	192	957	373.0	5.240000e-137	93.299	181	11	191	2	2	1042311
3	A0A2T3YV90	G5EB26	890	886	876	28	887	39	882	2200	852.0	0.000000e+00	51.712	453	375	593	48	16	1042311
4	A0A2T3YR98	C8VNJ9	1009	580	52	19	69	466	517	136	57.0	4.750000e-11	50.000	26	25	33	1	1	1042311
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
307	A0A2T3ZZ23	Q5BBS5	145	172	133	13	144	26	158	366	145.0	3.450000e-48	54.887	73	59	95	1	1	983964
308	A0A2T4AVC2	Q5AW17	319	577	60	139	198	158	214	123	52.0	2.440000e-10	50.000	30	27	38	3	1	983964
309	A0A2T4AQJ4	C8VT14	1321	1313	1240	71	1294	80	1291	3377	1305.0	0.000000e+00	54.032	670	526	877	44	16	983964
310	A0A2T4A759	C8VJ72	518	485	481	27	477	30	476	919	358.0	8.440000e-122	50.936	245	172	303	64	23	983964

Se descartaron a través del filtro alrededor de 586 filas correspondientes a porcentajes de identidad menores de 50% al realizar el alineamiento entre el query y el subject.

Para la resolución del problema sobre el match de proteínas de qacc y sacc, en donde tenemos varias (del qacc) con % de identidad diferentes que hacen match con una sola del sacc, se realizó la búsqueda por **Phobius** y **Pfam** de los dominios, familias y regiones de las proteínas (**Anexo 2)**, con la finalidad de que más adelante, se pudieran identificar las proteínas con mayor similitud dada su arquitectura proteica y se descartaran aquellas que no cumplieran. Al finalizar esta búsqueda, se concatenaron ambos (Pfam y Phobius) por medio del comando: metadata=pd.concat([pfam,fobius]).drop_duplicates().sort_values(by='start',ascending=True).reset_index(drop=True). **Tabla 5.**

Tabla 5. Creación del "metadata".

```
1 metadata
```

	Entry	start	end	name	type
0	L0PER5	1	134	Ras	Domain
1	Q5AUJ7	1	39	sig_p	Signal_peptide
2	C8VUM1	1	25	sig_p	Signal_peptide
3	Q9P4U6	1	25	sig_p	Signal_peptide
4	T0KIB8	1	230	Pkinase	Domain
...	...	...	...	...	...
5908	G9NBT7	996	1112	Response_reg	Domain
5910	A0A2T3Z679	997	1113	Response_reg	Domain
5911	A0A2T3Z765	997	1148	HATPase_c	Domain
5912	G2QAG0	999	1115	Response_reg	Domain
5913	A0A2T3ZW12	999	1115	Response_reg	Domain

5421 rows × 5 columns

La columna "Entry" guarda los valores de los id de cada proteína, "start" y "end" indican el inicio y fin de la región en donde se encuentra el dominio, péptido señal, en "name" y "type" se describe el nombre del dominio, péptido señal.

A continuación, las proteínas potencialmente ortólogas fueron identificadas utilizando un umbral de 0.5 con base al **índice de jaccard** (estadística para medir las similitudes entre dos conjuntos), en este caso se tomaron los datos del metadata y table1: qacc y sacc). Aquellas proteínas con un número de dominios igual al valor esperado (subject) fueron consideradas, las proteínas con un valor diferente fueron descartadas, esto asegura que ambas proteínas tengan la misma arquitectura (en ella comparten dominios, regiones, longitudes) **(Figura 18).**

Se consideró que el valor de 0 corresponde a aquellas que no comparten ningún dominio, y >=0.5 a aquellas que comparten un número de dominios igual al valor esperado y que estos son idénticos o casi idénticos. (**Figura 19).**

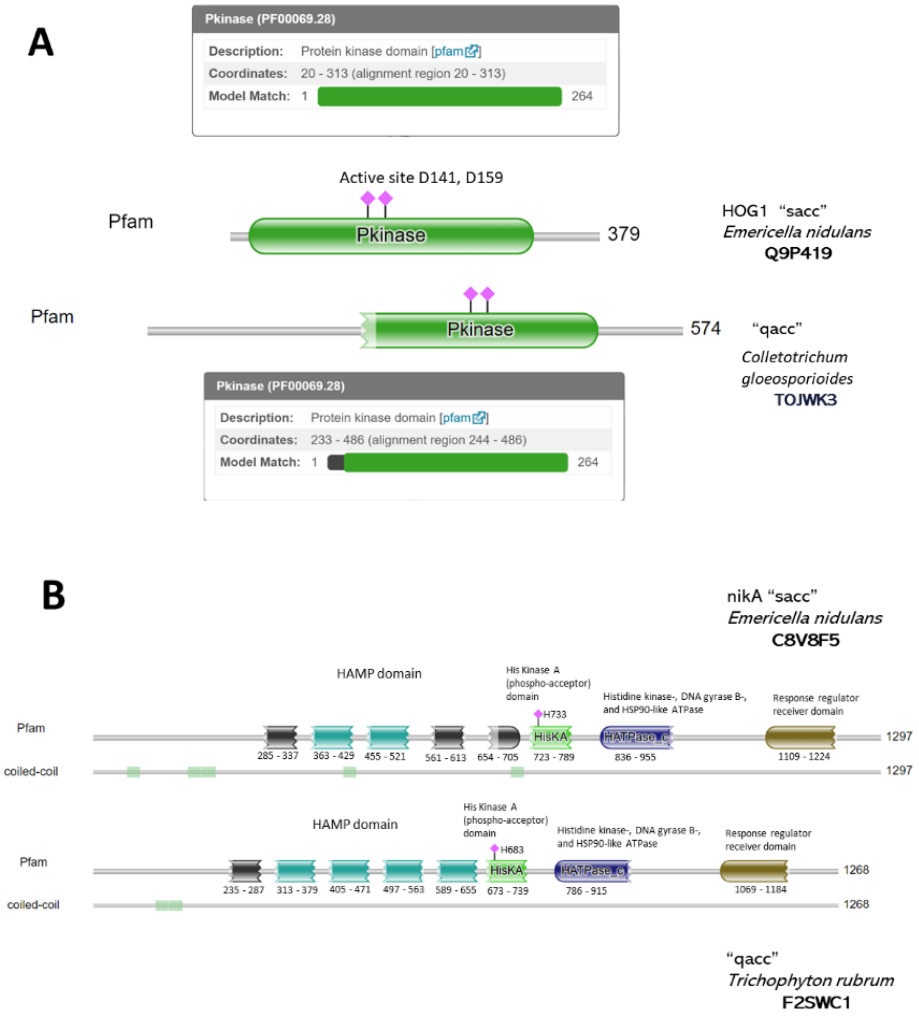

Figura 18. Comparación de arquitecturas. Realizada a partir de Pfam (**https://www.ebi.ac.uk/interpro/entry/pfam/#table**) y Hmmer (**https://www.ebi.ac.uk/Tools/hmmer/**).

A. Comparación entre la proteína Hog1 de *E. nidulans* y la proteína potencialmente ortóloga TOJWK3 de *C. gloeosporioides*: ambas contienen un solo dominio "Protein Kinase" y dos sitios activos, aunque difieren en la longitud de la proteína y por lo tanto también en las posiciones de los sitios activos y el dominio. **B.** Comparación de la proteína nikA de *E. nidulans* y la proteína potencialmente ortóloga F2SWC1 de *T. rubrum*: en este caso no solo tienen un dominio, poseen

varios dominios “HAMP” (los cuales conectan los dominios sensoriales extracelulares con los de señalización intracelular), el dominio “His Kinase A (phospho-acceptor)” con sitio activo, un dominio ATPasa “Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase”, además del dominio del receptor del regulador de respuesta, aunque difieren en las posiciones y longitud de la proteína, ambos comparten el mismo número de dominios (los cuales son los mismos) y el mismo orden como en el caso de **A**. En ambos casos se cumple que “qacc=índice de jaccard=sacc” ya que comparten el mismo número de dominios.

```
def jaccard_similarity(s1="qacc", s2="sacc"):
    return float(len(s1.intersection(s2)) / len(s1.union(s2)))
```

```
dominio=[]
for i,j in zip(table1.qacc,table1.sacc):
    dominio.append([metadata[metadata.Entry==i],metadata[metadata.Entry==j]])
```

```
dom2=[]
dom3=[]
for q,s in dominio:
    q=q.sort_values(by ='start',ascending=True).reset_index(drop=True)
    s=s.sort_values(by ='start',ascending=True).reset_index(drop=True)
    d=[]
    for a,b in zip(q.name,s.name):
        c1=set(a)
        c2=set(b)
        if jaccard_similarity(s1=c1, s2=c2)>=0.5:
            #print(q.entry.tolist()[0],s.entry.tolist()[0],a,b,jaccard_similarity(s1=c1, s2=c2))
            d.append([a,b])
    if len(q)==len(d)==len(s):
        #print(q,s)
        dom2.append([q,s])
        dom3.append([q.Entry.tolist()[0],s.Entry.tolist()[0]])
```

Figura 19. Código para la identificación de proteínas potencialmente ortólogas con base al índice de jaccard.

Se llamó a la función jaccard_similarity y se asignaron los valores del “qacc” y “sacc” para medir su similaridad, dado el tamaño de la intersección de ambos conjuntos divido por el tamaño de la unión de ambos. Así la intersección da el número de datos compartidos entre ambos conjuntos y la unión da el número total de datos (compartidos y no compartidos) en ambos conjuntos. Se realizó un zip de “table1: qacc y sacc” y se conjunto en el “metadata”. Se utilizó un bucle “for loop” para ordenar los valores de forma ascendente y se prosiguió a ejecutar el índice de jaccard con un umbral de >=0.5, si los valores correspondieron (q=d=s) entonces se conjuntaron los datos “q” y “s”.

Los datos de dom3: “qacc” () y “sacc” () fueron guardados en la variable “dom4” y por medio del método ***. merge***, se unió estos datos con los de table1 y se guardó en la variable “col”. comando: dom4=DataFrame(dom3,columns=["qacc","sacc"]), col=dom4.merge(table1,on=["qacc","sacc"],how="left").

Se creó un bucle “for loop” con el fin de devolver los valores únicos de “sacc” y como segunda acción ordenó los valores del % de identidad (pident) de forma descendente. Se guardó en la variable final1 y se concatenó, se generó el DataFrame “final2”. **Tabla 6.**

Tabla 6. DataFrame “final2”. Vista previa; dimensión real: 191 filas por 20 columnas.

In [60]:
```
final1=[]
for i in col.sacc.unique():
    df=col[col.sacc ==i].sort_values(by ='pident',ascending=False).reset_index(drop=True)
    #colduplicate=df.drop_duplicates(subset = 'organism',  keep = 'first')
    final1.append(df)

```

In [61]:
```
final2= pd.concat(final1)
final2
```

Out[61]:

	qacc	sacc	qlen	slen	length	qstart	qend	sstart	send	score	bitscore	evalue	pident	nident	mismatch	positive	gaps	gapopen
0	Q5BBS5	Q5BBS5	172	172	172	1	172	1	172	889	347.0	2.110000e-127	100.000	172	0	172	0	0
1	M7UZ28	Q5BBS5	190	172	132	57	188	23	153	448	177.0	5.330000e-60	65.152	86	45	109	1	1
2	F9X0A0	Q5BBS5	143	172	135	1	135	23	156	403	159.0	7.810000e-54	59.259	80	54	102	1	1
3	A0A061HKN7	Q5BBS5	273	172	147	16	161	17	162	423	167.0	5.070000e-55	57.143	84	61	108	2	2
4	F2SX35	Q5BBS5	173	172	170	5	172	10	159	431	170.0	1.090000e-57	57.059	97	51	115	22	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	C8VNJ9	C8VNJ9	580	580	580	1	580	1	580	3098	1197.0	0.000000e+00	100.000	580	0	580	0	0
0	Q5AXD9	Q5AXD9	821	821	821	1	821	1	821	4001	1545.0	0.000000e+00	100.000	821	0	821	0	0
0	C8VJ72	C8VJ72	485	485	485	1	485	1	485	2544	984.0	0.000000e+00	100.000	485	0	485	0	0

Los datos muestran un orden descendente (mayor a menor) con respecto al % de identidad ubicado en la columna “pident” y las proteínas del “sacc”, se observa el mejor match entre las proteínas del hongo *E. nidulans* y distintos hongos filamentosos como *Botrytis cinerea, Zymoseptoria tritici*, *Blumeria graminis*, *Tricophyton rubrum,* por mencionar algunos.

La tabla “final2” se guardó en formato fasta “final2_Larissa.fasta” y posteriormente se realizó la nueva base de datos **BLAST** con los datos del fasta, ejecutando el comando “subprocess.call('makeblastdb -in final2_Larissa.fasta -dbtype prot -parse_seqids -out DB2/final2_Larissa', shell = True)”, esto con el fin de poder realizar el enfoque Blastp All vs All en donde se realiza la comparación de todo contra todo y con ello se agrupan aquellas proteínas con similaridades altas.

5.4.1 BLASTP All vs All

Ejecuta una búsqueda BLASTP reciproca de todos contra todos para buscar similitudes de proteínas potencialmente ortólogas ante las proteínas del hongo de referencia: *E. nidulans* y compararse también entre sí. **Figura 20.**

```
"subprocess.call('blastp -db DB2/final2_Larissa -query final2_Larissa.fasta -evalue 1E-6 -outfmt "6
qacc sacc qlen slen length qstart qend sstart send score bitscore evalue pident nident mismatch
positive gaps gapopen" -max_target_seqs '+str(len(iden_gen))+' -max_hsps 1 -out
all_vs_all_Larissa.txt', shell = True)"
```

Figura 20. BLASTP ALL_VS_ALL.

max_target_seqs consideró en este caso los datos de "iden_gen" que corresponden a las 21 proteínas implicadas en la respuesta a estrés osmótico y sus respectivos identificadores. max_hsps 1 tomó el mejor hit (la mejor proteína dentro del query) que tiene una mayor similitud con respecto a la proteína del subject, es decir, si hay 21 proteínas implicadas en la respuesta a estrés osmótico, se esperaría encontrar el mejor hit para cada una (en una situación ideal se encontrarían las 21 proteínas potencialmente ortólogas para cada hongo filamentoso).

Al ejecutar BLASTP se quedaron solo los % de identidad más altos de cada proteína y los mejores de cada comparación (referente a cada comparación entre cada proteína vs proteína). Por ejemplo: B con D= 90% y B con D=85%, se tomó el de mayor %, en este caso el primero con un 90% de identidad.

Se consideró la longitud del alineamiento (columna "length") y el porcentaje de identidad (columna "pident"), como principales parámetros para la identificación de proteínas potencialmente ortólogas. La longitud de alineamiento indica que hay regiones donde son similares las secuencias ("query", "subject"), por lo que su valor debe ser allegado a la longitud del "subject" (columna "slen") y el porcentaje de identidad representa la conservación entre las secuencias, es decir, que tan idénticas son.

El valor **E** se tomó como una indicación de la significancia estadística de la alineación por pares dada y el reflejo del tamaño de la base de datos. Cuanto menor sea el valor E, más significativo será el impacto (Madden, 2010).

El resultado fue guardado en la variable "all_vs_all", **Tabla 7.**

Tabla 7. All vs All. Vista previa, dimensiones: 3345 filas × 20 columnas.

```
In [80]:  1 all_vs_all
```

	sacc	qlen	slen	length	qstart	qend	sstart	send	score	bitscore	evalue	pident	nident	mismatch	positive	gaps	gapopen	organism_sacc	organism_qacc
	Q5BBS5	172	172	172	1	172	1	172	889	347.0	1.620000e-126	100.000	172	0	172	0	0	Emericella nidulans Q5BBS5 (ypdA)	Emericella nidulans Q5BBS5 (ypdA)
	F2SX35	172	173	171	9	159	4	172	433	171.0	4.670000e-57	56.725	97	52	115	22	4	Trichophyton rubrum F2SX35	Emericella nidulans Q5BBS5 (ypdA)
	)A061HKN7	172	273	147	17	162	16	161	425	168.0	1.450000e-54	57.143	84	61	108	2	2	Blumeria graminis f. sp. tritici A0A061HKN7	Emericella nidulans Q5BBS5 (ypdA)
	T0LW75	172	150	135	20	153	12	145	379	150.0	3.580000e-49	57.037	77	56	99	2	2	Colletotrichum gloeosporioides T0LW75	Emericella nidulans Q5BBS5 (ypdA)
	0A2T3ZZ23	172	145	133	26	158	13	144	366	145.0	2.750000e-47	54.887	73	59	95	1	1	Trichoderma harzianum (Hypocrea lixii) A0A2T3ZZ23	Emericella nidulans Q5BBS5 (ypdA)

Columnas:'qacc','sacc','qlen','slen','length','qstart','qend','sstart','send','score','bitscore','evalue','pident','nident','mismatch','positive','gaps','gapopen','organism_sacc','organism_qacc'. Se observa algunas comparaciones entre la proteína "ypdA" de *E. nidulans* y "ypdA" de *E. nidulans*, la proteína "ypdA" de *E. nidulans* y las proteínas potencialmente ortólogas de los hongos filamentosos como: *T. rubrum*, *B. graminis f. sp. tritici*, *C. gloeosporioides*, *T. harzium*, y estos entre sí, por mencionar algunos.

Para saber cuántas proteínas se obtuvieron, se ejecutó el comando "len(all_vs_all.sacc.unique().tolist()), len(all_vs_all.qacc.unique().tolist())", el resultado fue de (191, 191) proteínas obtenidas, es razonable debido a los filtros realizados (si todas las proteínas cumplieran los datos serían: 21 proteínas x 17 hongos filamentosos= 357 proteínas obtenidas). Estas 191 proteínas cumplieron con los parámetros: % de identidad >=50%, índice de jaccard con un umbral de >=0.5 y el mejor hit para cada proteína. Es por lo que han sido consideradas como proteínas potencialmente ortólogas a la vía de señalización de respuesta a estrés osmótico, a partir del hongo *E. nidulans*.

5.4.2 Análisis de Clustering

El análisis de agrupamiento (Clustering) es el nombre dado a un conjunto de técnicas cuyo objetivo es agrupar un set de objetos de datos en clústeres (similar entre sí dentro del mismo clúster/ diferente a los objetos de otros clústeres) (Chuan *et al.,* 2006).

Se realizó un análisis de clustering usando la métrica de correlación ya que esta permitió identificar más grupos de proteínas ortólogas en múltiples organismos.

Para ello los datos de "All vs All" se procesaron a partir de all_vs_all[['organism_qacc', 'pident', 'qacc', 'sacc']] para llegar a una matriz de datos (**Tabla 8**).

Tabla 8. Matriz “matmat”. Vista previa, dimensiones: 191 filas x 191 columnas.

```
In [93]:  1 matmat = matrix.drop(columns = ['sacc', 'organism_sacc'])
          2 matmat
```

organism_qacc	Acremonium chrysogenum A0A086SUG6	Acremonium chrysogenum A0A086SXF1	Acremonium chrysogenum A0A086SY72	Acremonium chrysogenum A0A086SYL4	Acremonium chrysogenum A0A086T736	Acremonium chrysogenum A0A086TCN0	Acremonium chrysogenum A0A086TDN7	Acremonium chrysogenum A0A086TET7	Acremonium chrysogenum A0A086TFF9	Acremonium chrysogenum A0A086TFH5	...	Zym
0	0.000	0.000	0.00	0.000	0.0	0.0	0.000	0.000	0.000	0.0	...	
1	0.000	0.000	0.00	0.000	0.0	0.0	0.000	0.000	0.000	0.0	...	
2	0.000	32.012	0.00	0.000	0.0	0.0	0.000	0.000	61.873	0.0	...	
3	0.000	0.000	0.00	51.242	0.0	0.0	0.000	0.000	0.000	0.0	...	
4	91.579	0.000	0.00	0.000	0.0	0.0	0.000	0.000	0.000	0.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
186	94.819	0.000	0.00	0.000	0.0	0.0	0.000	0.000	0.000	0.0	...	
187	0.000	0.000	0.00	0.000	0.0	0.0	64.736	0.000	0.000	0.0	...	
188	0.000	0.000	0.00	0.000	0.0	0.0	0.000	0.000	0.000	0.0	...	
189	0.000	0.000	0.00	0.000	0.0	0.0	0.000	26.829	0.000	0.0	...	

A continuación, se representaron los datos en un dendograma (**Figura 21**) en cual se formaron 15 clústeres de proteínas potencialmente ortólogas entre de *E. nidulans* y los hongos filamentosos seleccionados. Estos clústeres se obtuvieron en función de la similaridad de las secuencias de aminoácidos y/o de la estructura secundaria de las proteínas. Para las proteínas del hongo modelo “*E. nidulans*”: msnA, atfA, msbA y Opy2, no se encontraron proteínas que cumplieran con los parámetros establecidos para la identificación de proteínas potencialmente ortólogas.

En el **Anexo 3** se encuentra el código utilizado para la ejecución del análisis y su representación (dendograma).

Por otro lado, en el clúster formado por las proteínas MAPK Hog1 (SakA) y su parálogo mpkC de *E. nidulans* (color morado) se visualizan a la misma distancia y muestran sus ortólogos para diferentes hongos filamentosos: *Myceliophthora thermophila* G2QKS1, *Acremonium chrysogenum* A0A086TCN0, *Trichoderma atroviride* G9P1X3 (*tmk3*), *Trichoderma asperellum* A0A2T3ZLL6, *Trichoderma reesei* G0R9Y0 (*tmk3*), *Trichoderma virens* G9MW46 (*tmk3*), *Trichoderma harzianum* A0A2T4AVG6, A0A2T4AS65, *Sporothrix schenckii* A0A0F2MGH6, *Lomentospora prolificans* A0A2N3N9Q1, *Zymoseptoria tritici* Q1KTF2 (HOG1_ZYMTI), *Botryotinia fuckeliana* M7TLS3 (M7TLS3_BOTF1), *Trichophyton rubrum* F2SP73, A0A080WMI3, A0A080WG74.

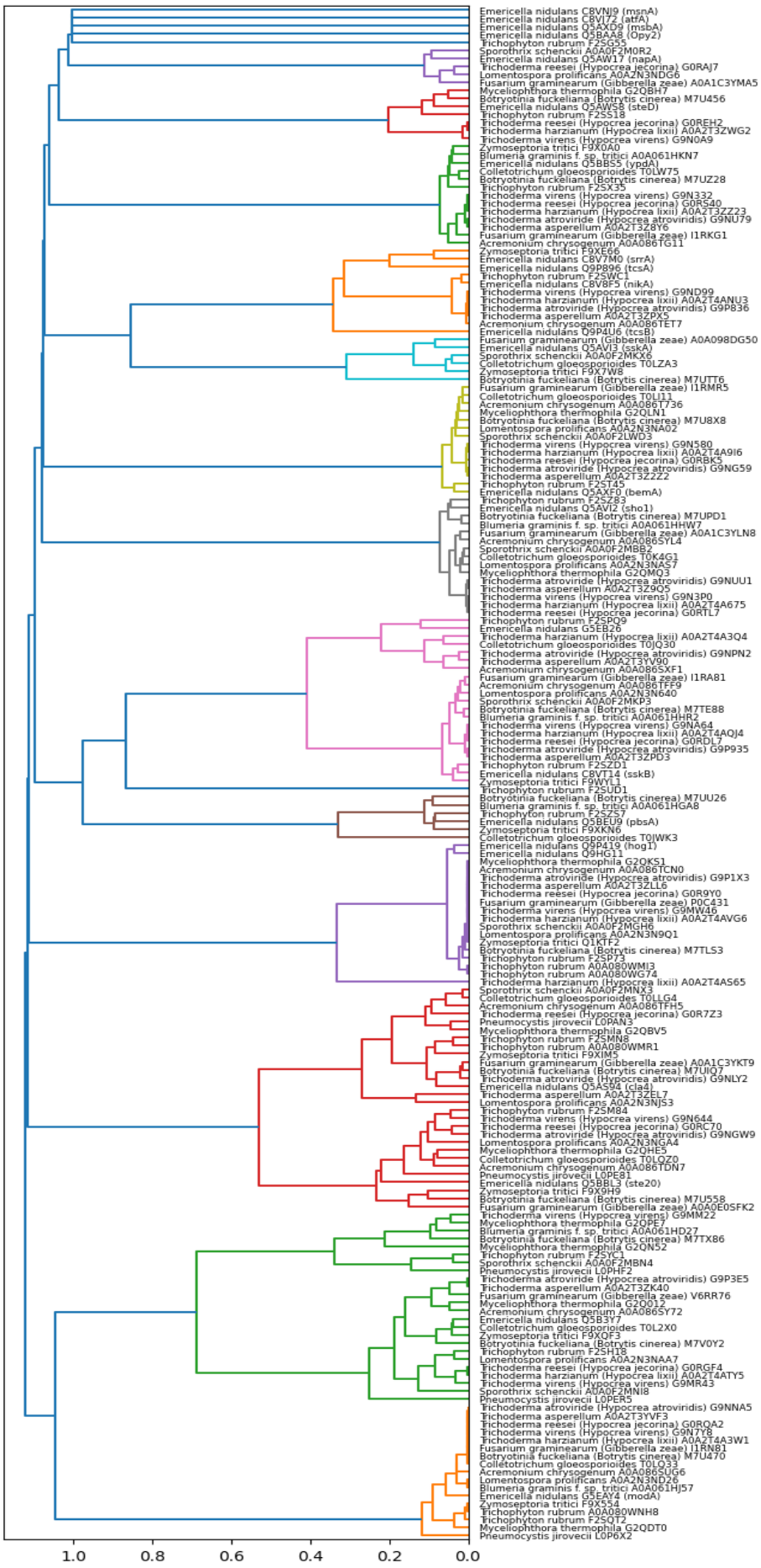
Emericella nidulans C8VNJ9 (msnA)
Emericella nidulans C8VI72 (atfA)
Emericella nidulans Q5AXD9 (msbA)
Emericella nidulans Q5BAA8 (Opy2)
Trichophyton rubrum F2SQS5
Sporothrix schenckii A0A0F2M0R2
Emericella nidulans Q5AW17 (napA)
Trichoderma reesei (Hypocrea jecorina) G0RAJ7
Lomentospora prolificans A0A2N3NDG6
Fusarium graminearum (Gibberella zeae) A0A1C3YMA5
Myceliophthora thermophila G2QBH7
Botryotinia fuckeliana (Botrytis cinerea) M7U456
Emericella nidulans Q5AWS8 (steD)
Trichophyton rubrum F2SS18
Trichoderma reesei (Hypocrea jecorina) G0REH2
Trichoderma harzianum (Hypocrea lixii) A0A2T3ZWG2
Trichoderma virens (Hypocrea virens) G9N0A9
Zymoseptoria tritici F9X0A0
Blumeria graminis f. sp. tritici A0A061HKN7
Emericella nidulans Q5BBS5 (ypdA)
Colletotrichum gloeosporioides T0LW75
Botryotinia fuckeliana (Botrytis cinerea) M7UZ28
Trichophyton rubrum F2SX35
Trichoderma virens (Hypocrea virens) G9N332
Trichoderma reesei (Hypocrea jecorina) G0RS40
Trichoderma harzianum (Hypocrea lixii) A0A2T3ZZ23
Trichoderma atroviride (Hypocrea atroviridis) G9NU79
Trichoderma asperellum A0A2T3Z8Y6
Fusarium graminearum (Gibberella zeae) I1RKG1
Acremonium chrysogenum A0A086TG11
Zymoseptoria tritici F9XE66
Emericella nidulans C8V7M0 (srrA)
Emericella nidulans Q9P896 (tcsA)
Trichophyton rubrum F2SWC1
Emericella nidulans C8V8F5 (nikA)
Trichoderma virens (Hypocrea virens) G9ND99
Trichoderma harzianum (Hypocrea lixii) A0A2T4ANU3
Trichoderma atroviride (Hypocrea atroviridis) G9P836
Trichoderma asperellum A0A2T3ZPX5
Acremonium chrysogenum A0A086TET7
Emericella nidulans Q9P4U6 (tcsB)
Fusarium graminearum (Gibberella zeae) A0A098DG50
Emericella nidulans Q5AVI3 (sskA)
Sporothrix schenckii A0A0F2MKX6
Colletotrichum gloeosporioides T0LZA3
Zymoseptoria tritici F9X7W8
Botryotinia fuckeliana (Botrytis cinerea) M7UTT6
Fusarium graminearum (Gibberella zeae) I1RMR5
Colletotrichum gloeosporioides T0LJ11
Acremonium chrysogenum A0A086T736
Myceliophthora thermophila G2QLN1
Botryotinia fuckeliana (Botrytis cinerea) M7U8X8
Lomentospora prolificans A0A2N3NA02
Sporothrix schenckii A0A0F2LWD3
Trichoderma virens (Hypocrea virens) G9N580
Trichoderma harzianum (Hypocrea lixii) A0A2T4A9I6
Trichoderma reesei (Hypocrea jecorina) G0RBK5
Trichoderma atroviride (Hypocrea atroviridis) G9NG59
Trichoderma asperellum A0A2T3Z2Z2
Trichophyton rubrum F2ST45
Emericella nidulans Q5AXF0 (bemA)
Trichophyton rubrum F2SZ83
Emericella nidulans Q5AVI2 (sho1)
Botryotinia fuckeliana (Botrytis cinerea) M7UPD1
Blumeria graminis f. sp. tritici A0A061HHW7
Fusarium graminearum (Gibberella zeae) A0A1C3YLN8
Acremonium chrysogenum A0A086SYL4
Sporothrix schenckii A0A0F2MB82
Colletotrichum gloeosporioides T0K4G1
Lomentospora prolificans A0A2N3NAS7
Myceliophthora thermophila G2QMQ3
Trichoderma atroviride (Hypocrea atroviridis) G9NUU1
Trichoderma asperellum A0A2T3Z9Q5
Trichoderma virens (Hypocrea virens) G9N3P0
Trichoderma harzianum (Hypocrea lixii) A0A2T4A675
Trichoderma reesei (Hypocrea jecorina) G0RTL7
Trichophyton rubrum F2SPQ9
Emericella nidulans G5EB26
Trichoderma harzianum (Hypocrea lixii) A0A2T4A3Q4
Colletotrichum gloeosporioides T0JQ30
Trichoderma atroviride (Hypocrea atroviridis) G9NPN2
Trichoderma asperellum A0A2T3YV90
Acremonium chrysogenum A0A086SXF1
Fusarium graminearum (Gibberella zeae) I1RA81
Acremonium chrysogenum A0A086TFF9
Lomentospora prolificans A0A2N3N640
Sporothrix schenckii A0A0F2MKP3
Botryotinia fuckeliana (Botrytis cinerea) M7TE88
Blumeria graminis f. sp. tritici A0A061HHR2
Trichoderma virens (Hypocrea virens) G9NA64
Trichoderma harzianum (Hypocrea lixii) A0A2T4AQJ4
Trichoderma reesei (Hypocrea jecorina) G0RDL7
Trichoderma atroviride (Hypocrea atroviridis) G9P935
Trichoderma asperellum A0A2T3ZPD3
Trichophyton rubrum F2SZD1
Emericella nidulans C8VT14 (sskB)
Zymoseptoria tritici F9WYL1
Trichophyton rubrum F2SUD1
Botryotinia fuckeliana (Botrytis cinerea) M7UU26
Blumeria graminis f. sp. tritici A0A061HGA8
Trichophyton rubrum F2SZS7
Emericella nidulans Q5BEU9 (pbsA)
Zymoseptoria tritici F9XKN6
Colletotrichum gloeosporioides T0JWK3
Emericella nidulans Q9P419 (hog1)
Emericella nidulans Q9HG11
Myceliophthora thermophila G2QKS1
Acremonium chrysogenum A0A086TCN0
Trichoderma atroviride (Hypocrea atroviridis) G9P1X3
Trichoderma asperellum A0A2T3ZLL6
Trichoderma reesei (Hypocrea jecorina) G0R9Y0
Fusarium graminearum (Gibberella zeae) P0C431
Trichoderma virens (Hypocrea virens) G9MW46
Trichoderma harzianum (Hypocrea lixii) A0A2T4AVG6
Sporothrix schenckii A0A0F2MGH6
Lomentospora prolificans A0A2N3N9Q1
Zymoseptoria tritici Q1KTF2
Botryotinia fuckeliana (Botrytis cinerea) M7TLS3
Trichophyton rubrum F2SP73
Trichophyton rubrum A0A080WMJ3
Trichophyton rubrum A0A080WG74
Trichoderma harzianum (Hypocrea lixii) A0A2T4AS65
Sporothrix schenckii A0A0F2MNX3
Colletotrichum gloeosporioides T0LLG4
Acremonium chrysogenum A0A086TFH5
Trichoderma reesei (Hypocrea jecorina) G0R7Z3
Pneumocystis jirovecii L0PAN3
Myceliophthora thermophila G2QBV5
Trichophyton rubrum F2SMN8
Trichophyton rubrum A0A080WMR1
Zymoseptoria tritici F9XIM5
Fusarium graminearum (Gibberella zeae) A0A1C3YKT9
Botryotinia fuckeliana (Botrytis cinerea) M7UIQ7
Trichoderma atroviride (Hypocrea atroviridis) G9NLY2
Emericella nidulans Q5AS94 (cla4)
Trichoderma asperellum A0A2T3ZEL7
Lomentospora prolificans A0A2N3NJS3
Trichophyton rubrum F2SM84
Trichoderma virens (Hypocrea virens) G9N644
Trichoderma reesei (Hypocrea jecorina) G0RC70
Trichoderma atroviride (Hypocrea atroviridis) G9NGW9
Lomentospora prolificans A0A2N3NGA4
Myceliophthora thermophila G2QHE5
Colletotrichum gloeosporioides T0LQZ0
Acremonium chrysogenum A0A086TDN7
Pneumocystis jirovecii L0PE81
Emericella nidulans Q5BBL3 (ste20)
Zymoseptoria tritici F9X9H9
Botryotinia fuckeliana (Botrytis cinerea) M7U558
Fusarium graminearum (Gibberella zeae) A0A0E0SFK2
Trichoderma virens (Hypocrea virens) G9MM22
Myceliophthora thermophila G2QPE7
Blumeria graminis f. sp. tritici A0A061HD27
Botryotinia fuckeliana (Botrytis cinerea) M7TX86
Myceliophthora thermophila G2QN52
Trichophyton rubrum F2SYC1
Sporothrix schenckii A0A0F2MBN4
Pneumocystis jirovecii L0PHF2
Trichoderma atroviride (Hypocrea atroviridis) G9P3E5
Trichoderma asperellum A0A2T3ZK40
Fusarium graminearum (Gibberella zeae) V6RR76
Myceliophthora thermophila G2Q932
Acremonium chrysogenum A0A086SY72
Emericella nidulans Q5B3Y7
Colletotrichum gloeosporioides T0L2X0
Zymoseptoria tritici F9XQF3
Botryotinia fuckeliana (Botrytis cinerea) M7V0Y2
Trichophyton rubrum F2SH38
Lomentospora prolificans A0A2N3NAA7
Trichoderma reesei (Hypocrea jecorina) G0RGF4
Trichoderma harzianum (Hypocrea lixii) A0A2T4ATY5
Trichoderma virens (Hypocrea virens) G9MR43
Sporothrix schenckii A0A0F2MNI8
Pneumocystis jirovecii L0PER5
Trichoderma atroviride (Hypocrea atroviridis) G9NNA5
Trichoderma asperellum A0A2T3YVF3
Trichoderma reesei (Hypocrea jecorina) G0RQA2
Trichoderma virens (Hypocrea virens) G9N7Y8
Trichoderma harzianum (Hypocrea lixii) A0A2T4A3W1
Fusarium graminearum (Gibberella zeae) I1RN81
Botryotinia fuckeliana (Botrytis cinerea) M7U470
Colletotrichum gloeosporioides T0LQ33
Acremonium chrysogenum A0A086SUG6
Lomentospora prolificans A0A2N3ND26
Blumeria graminis f. sp. tritici A0A061HJ57
Emericella nidulans G5EAY4 (modA)
Zymoseptoria tritici F9X554
Trichophyton rubrum A0A080WNH8
Trichophyton rubrum F2SQT2
Myceliophthora thermophila G2QDT0
Pneumocystis jirovecii L0P6X2
1.0
0.8
0.6
0.4
0.2
0.0

Figura 21. Agrupación de secuencias utilizando la métrica: correlación.

La escala es de 0-1, donde 0 es altamente similar y 1 es altamente disimilar (diferente). Por lo que, si la distancia es menor, estará más cercana a cero. La distancia máxima la marca la línea azul oscuro que une los clústeres (color naranja, verde claro, rojo, morado, café, fucsia, gris, verde oliva, azul claro, etc.).

Los 15 clústeres obtenidos se visualizan en diferentes colores, cada uno de ellos representa un grupo de proteínas (de los hongos filamentosos seleccionados) que son potencialmente ortólogas a alguna proteína de referencia (del hongo modelo *E. nidulans*). En varios de los clústeres formados existe una alta similaridad del grupo hacia una proteína de referencia que a otra (aunque se encuentre en el mismo clúster), esto se ve reflejado en las distancias, las más cortas (cercanas a cero) indican mayor similaridad: Por ejemplo: F9XE66 (*Zymoseptoria tritici*) tiene mayor similaridad hacia srrA de *E. nidulans* en color naranja, que hacia la proteína tcsA de *E. nidulans* e incluso la distancia aumenta hacia la proteína tcsB (la cual también se encuentra en el mismo clúster).

Los clústeres más similares fueron los de las especies de *Trichoderma*: *T. reesei, T. harzianum, T. virens, T. atroviride y T. asperellum* dada su selección para generar un contraste entre especies de un mismo género. Recordemos que estos hongos se han adaptado a distintas regiones geográficas en el mundo (son cosmopolitas), debido a que comparten ciertas características generales: son oportunistas de plantas y de humanos inmunodeprimidos, pueden alimentarse de material celulolítico y pueden establecerse en el suelo y colonizar la rizosfera, por lo que se puede deducir que varias de las proteínas que participan en vías de señalización pueden ser similares.

En la **Tabla 9**. se muestran las proteínas con mayor similaridad entre especies *Trichoderma* y *E. nidulans*, entre ellas se encuentran distintas proteínas quinasas (que participan en la transducción de señales), en las que se ha demostrado que hay un alto grado de conservación ante la evolución del género *Trichoderma* (Kubicek *et al.,* 2019).

Tabla 9. Proteínas identificadas en especies *Trichoderma.*

Trichoderma	steD (ste50)	*ypdA*	nikA	BemA (Bem1)	Sho1	sskB	hog1 (SakA)/ MpkC	cla 4	ste 20	Rac1	modA (cdc42)
reesei	*G0REH2*	*G0RS40*		G0RBK5	G0RTL7	G0RDL7	G0R9Y0	G0R7Z3	G0RC70	G0RGF4	G0RQA2
harzianum	*A0A2T3ZWG2*	*A0A2T3ZZ23*	A0A2T4ANU3	A0A2T4A9I6 scd2/ral3	A0A2T4A675	A0A2T4AQJ4	A0A2T4AVG6, A0A2T4AS65			A0A2T4ATY5	A0A2T4A3W1
virens	*G9N0A9*	*G9N332*	G9ND99	G9N580 scd2/ral3	G9N3P0	G9NA64	G9MW46		G9N644	G9MR43	G9N7Y8
atroviride		*G9NU79*	G9P836	G9NG59 scd2/ral3	G9NUU1	G9P935	G9P1X3	G9NLY2	G9NGW9	G9P3E5	G9NNA5
asperellum		*A0A2T3Z8Y6*	A0A2T3ZPX5	A0A2T3Z2Z2 scd2/ral3	A0A2T3Z9Q5	A0A2T3ZPD3	A0A2T3ZLL6	A0A2T3ZEL7		A0A2T3ZK40	A0A2T3YVF3

La tabla está organizada de acuerdo con las proteínas potencialmente ortólogas identificadas. Se muestra el identificador taxonómico UniProt de cada proteína correspondiente a cada especie *Trichoderma*. En *harzianum/virens* (pertenecen a un mismo clado) el número de proteínas identificadas fue casi el mismo difiriendo solo en la MAP4K ste20 (se une a BemA-Cdc42), lo mismo ocurrió para *atroviride/asperellum* (clado *Trichoderma*).

También se encontró una alta similaridad entre las proteínas de estas especies de *Trichoderma* y hongos de la misma clase "*Sordariomycetes*": *F. graminearum "*Fusarium Head Blight (FHB)", *A. chrysogenum* (productor de Cefalosporina), *L. prolificans* (lomentosporiosis en humanos), *M. thermophila* (enzimas y biomasa vegetal), *C. gloeosporioides* (hongo de la antracnosis), *S. schenckii* (esporotricosis en humanos) y otros como *T. rubrum* (tinea en humanos) y *B. fuckeliana* (enfermedad de moho gris). Algunos de ellos son patógenos de plantas, su interacción con *Trichoderma* se ha asociado a su relación evolutiva entre micoparasitos y entomopatógenos, ya que tienen una relación patógeno-planta-micoparasito y se ha demostrado que, a través de estas interacciones, ha habido un proceso de adaptación y evolución en las plantas-hongo (Baluška, F., & Mancuso, S., 2013) y la relación entre hongos patógenos de humanos (Rokas, 2022), por lo que se sugiere que todos ellos tuvieron una coevolución adaptativa.

5.4.3 *Heatmap*

Con la finalidad de visualizar el grado de conservación de las proteínas potencialmente ortólogas de todos los hongos filamentosos se construyó un *heatmap*.

El grado de conservación en las proteínas nos indica que tan similar es la vía de señalización de respuesta a estrés osmótico entre los hongos filamentosos ascomicetos.

En la creación del Heatmap simétrico se tomaron en cuenta los datos de "mat1"

Tabla 10. "mat1" Dimensiones 191 filas x 191 columnas.

```
In [102]: 1 mat1 = mat.drop(columns = ['organism_sacc'])
          2 mat1
```

02]:

	Pneumocystis jirovecii L0P6X2	Myceliophthora thermophila G2QDT0	Trichophyton rubrum F2SQT2	Trichophyton rubrum A0A080WNH8	Zymoseptoria tritici F9X554	Emericella nidulans G5EAY4 (modA)	Blumeria graminis f. sp. tritici A0A061HJ57	Lomentospora prolificans A0A2N3ND26	Acremonium chrysogenum A0A086SUG6	Colletotrichum gloeosporioides T0LQ33	...	Fusari graminear (Gibber ze A0A1C3YM
0	100.000	72.222	63.598	71.698	72.596	72.170	72.727	72.300	74.178	74.757	...	
1	72.222	100.000	77.679	88.325	92.021	92.021	88.144	89.848	90.863	87.755	...	
2	63.598	77.679	100.000	87.727	80.556	81.364	77.880	80.995	80.090	77.273	...	
3	71.698	88.325	87.727	100.000	92.063	92.746	88.947	92.268	91.237	88.083	...	
4	72.596	92.021	80.556	92.063	100.000	91.489	91.398	91.837	95.699	92.432	...	
...	...	...	...	...	...	...	...	...	...	...	...	
186	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	
187	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	
188	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	

La tabla está organizada por correlaciones de acuerdo con el análisis de clustering usando la métrica "correlación". Los datos corresponden a los % de identidad obtenidos en el blastp All vs All, considerando solo los mejores hits entre proteínas vs proteínas de hongos filamentosos.

Con los datos de "mat1" se creó el *heatmap* (**Figura 22**), en el cual sirvió para visualizar que tan idénticas son las proteínas potencialmente ortólogas obtenidas de cada hongo ascomiceto, entre sí y entre las proteínas de *E. nidulans.*

En el **Anexo 4**, se encuentra el código utilizado para la creación del *Heatmap.*

Las proteínas con mayor grado de conservación son aquellas en colores de amarillo opaco a negro. Estas son: las proteínas MAPK hog1(controla la regulación osmótica de la transcripción de genes diana) y MAPK mpkC, paráloga de hog1, aunque si bien son similares en secuencia (62% identidad), se ha demostrado que MAPK mpkC responde mejor al estrés oxidativo en comparación a hog1 (*E. nidulans),* por lo que no responden transcripcionalmente de la misma manera a diferentes estreses ambientales, sin embargo, en ausencia del gen *mpkC*, *hog1* se expresa aún más, una característica que comparten es que ambas participan en la regulación de la integridad de la pared celular de conidios (Garrido *et al.,* 2018), por lo que es probable sus proteínas ortólogas en los hongos estudiados presenten también esta diferencia. Otras proteínas identificadas importantes son las MAPKK pbsA (pbsB), MAPKKK SskB, RR SskA, HPt ypdA, HK NikA, Osmosensor sho1, BemA, PAK cla4, PAK ste20, RacA (Rho GTPasa), GTPasa modA (Cdc42).

Las proteínas msnA, atfA, msbA y Opy2 (en color azul oscuro ubicadas en la parte inferior izquierda de la **Figura 22)** que pertenecen a *E. nidulans* no se lograron identificar para los otros hongos filamentosos, lo que podría deberse a la diversificación de vías de señalización para género y especie en los hongos ascomicetos y al procesamiento de datos en este estudio.

Reafirmamos que, a nivel de proteína, la comparación de secuencias ortólogas permite predicciones sobre dominios funcionales putativos, mientras que a nivel de vía brinda la oportunidad de evaluar el nivel de conservación evolutiva de vías específicas (en este caso sobre la vía de estrés osmótico) y generar más adelante nuevas hipótesis para su análisis funcional.

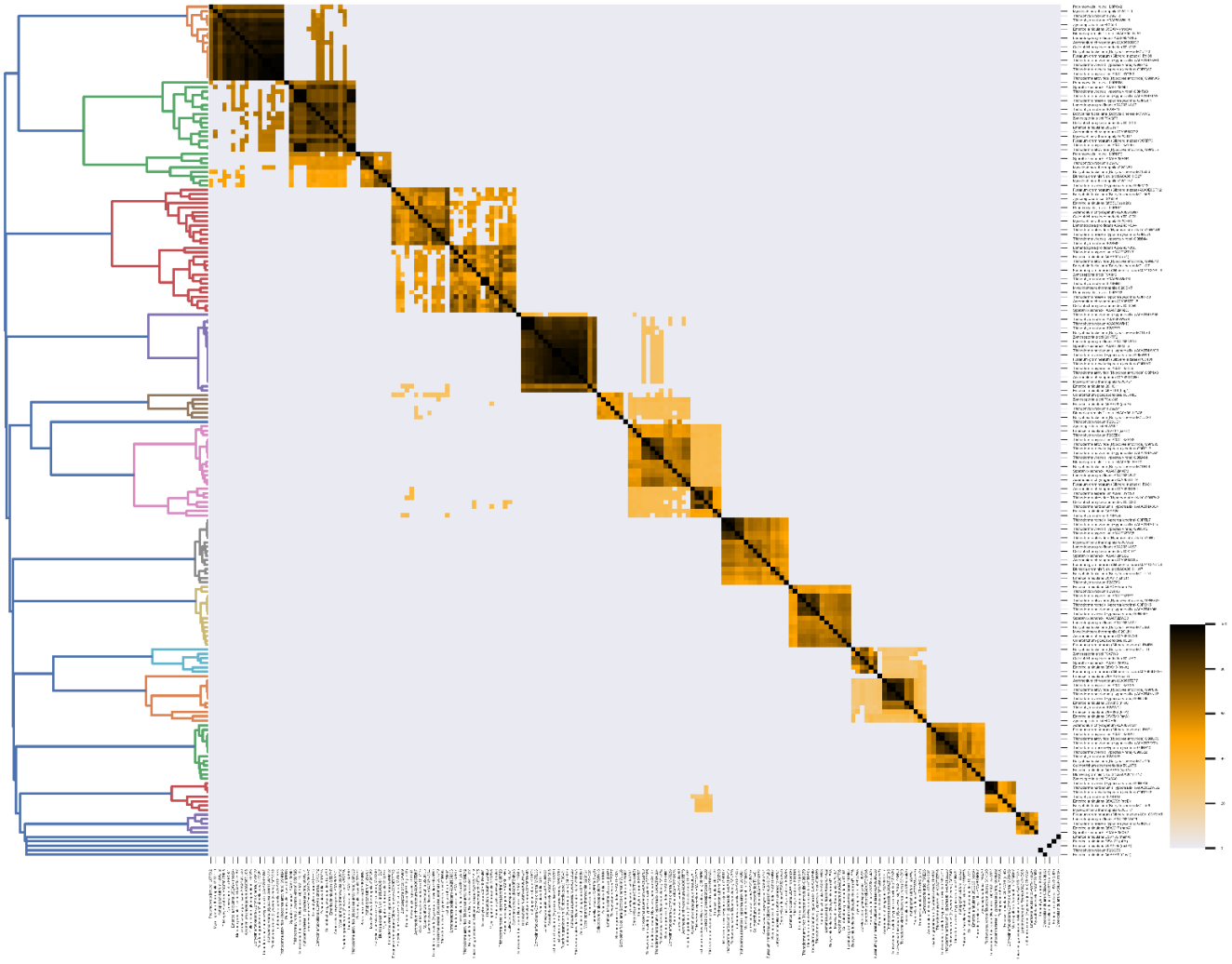

Figura 22. Identidad entre secuencias proteicas de *E. nidulans* y hongos filamentosos.

El *Heatmap* muestra el grado de conservación entre los 16 hongos filamentosos seleccionados en este trabajo y el hongo modelo *E. nidulans.* En el gráfico se observan agrupaciones en función de la similaridad de secuencias. La escala de color representa el % de identidad obtenido por Blastp All vs All usando un Evalue de 1E-6.

6. Objetivos y metas alcanzadas

- Para este estudio se seleccionaron hongos de gran interés industrial y hongos patógenos de plantas insectos y humanos que actualmente son relevantes en la comunidad científica.
- Utilizando la base de datos NCBI y UniProtKB y a través de un comando ejecutado en Jupyter Notebook se obtuvieron los proteomas de los hongos filamentosos seleccionados.
- La construcción de bases de datos individuales con los proteomas descargados se descartó para hacer una sola base de datos con las secuencias de las 21 proteínas de referencia, siendo esta una forma más práctica de realizar el Blastp.
- Mediante las bases de datos: UniProt, NCBI protein y AspGD, se identificaron las 21 proteínas que participan en la vía de señalización, así mismo a través de **https://www.ebi.ac.uk/interpro/** se logró visualizar y extraer su arquitectura. Se logró identificar las proteínas que participan en la vía de señalización de respuesta a estrés osmótico en especies de *Aspergillus*.
- En el programa Blastp se compararon las secuencias de los hongos seleccionados contra la base de datos creada a partir de *Aspergillus* y se realizaron varios filtros como: % de identidad >=50%, índice de jaccard con un umbral de >=0.5 y el mejor hit de cada proteína, para después poder identificar las proteínas ortólogas al ejecutar una búsqueda BLAST All vs All, dando como resultado 191 proteínas ortólogas
- Al realizar un análisis de agrupamiento (Clustering) con las proteínas obtenidas y utilizando la métrica de correlación, se obtuvo un dendograma en donde se pudo visualizar por colores los 15 clústeres formados y su similaridad a partir de una escala 0-1 en donde cero es alta similaridad.
- Al crear un Heatmap a partir de los datos de BLAST All vs All y las correlaciones del análisis de clustering, se obtuvo una buena visualización del grado de conservación de las proteínas y por tanto la similaridad de la vía entre hongos *Ascomycota*.

7. Conclusiones

- Se identificaron las proteínas implicadas en la vía de señalización a estrés osmótico: Sistema de dos componentes: HK, HPt y RR y la cascada HOG MAPK: MAPKKK, MAPKK y MAPK caracterizadas en el hongo modelo *Aspergillus nidulans/ Emericella nidulans*.
- Se identificaron un total de 191 proteínas potencialmente ortólogas implicadas en la vía de señalización de respuesta a estrés osmótico en hongos filamentosos pertenecientes a la división *Ascomycota* de gran interés industrial, científico, médico, farmacéutico y agrícola.
- Tras el Análisis de Clustering en el que se formaron 15 clústeres de proteínas con alta similaridad, algunos grupos se integraron más de 2 proteínas del hongo modelo *E. nidulans,* dando a resaltar la conservación de arquitecturas y su función de dominio similar.
- El género *Trichoderma* ha destacado en el estudio *in silico* debido a la gran conservación de arquitectura de dominio en proteínas quinasas.
- La interfaz Jupyter notebook y el lenguaje informático Python son herramientas útiles para jóvenes investigadores y han favorecido el estudio *in sillico* de proteínas.

8. Referencias

- Adnan, M., Islam, W., Shabbir, A., Khan, K. A., Ghramh, H. A., Huang, Z., Chen, H. Y. H., & Lu, G. D. (2019). Plant defense against fungal pathogens by antagonistic fungi with Trichoderma in focus. *Microbial Pathogenesis*, 129, 7–18.
- Aghcheh, R.K. & Braus, G.H. (2018). Importance of Stress Response Mechanisms in Filamentous Fungi for Agriculture and Industry. In: Skoneczny, M. (eds) Stress Response Mechanisms in Fungi. Springer, Cham. **https://doi.org/10.1007/978-3-030-00683-9_6.**
- Bagewadi, Z. K., Mulla, S. I., & Ninnekar, H. Z. (2018). Response surface methodology-based optimization of keratinase production from *Trichoderma harzianum* isolate HZN12 using chicken feather waste and its application in dehairing of hide. *Journal of Environmental Chemical Engineering*, 6(4), 4828–4839.
- Bahn, Y. S. (2008). Master and commander in fungal pathogens: the two-component system and the HOG signaling pathway. *Eukaryotic Cell*, 7(12), 2017–2036.
- Baluška, F., & Mancuso, S. (2013). Microorganism and filamentous fungi drive evolution of plant synapses. *Frontiers in Cellular and Infection Microbiology*, 3(44), 1-10.
- Barros, M. B., de Almeida Paes, R., & Schubach, A. O. (2011). *Sporothrix schenckii* and Sporotrichosis. *Clinical Microbiology Reviews*, 24(4), 633–654.
- Bayat, A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ (Clinical Research Ed.)*, 324(7344), 1018–1022.
- Bayram, Ö., Bayram, Ö. S., Ahmed, Y. L., Maruyama, J., Valerius, O., Rizzoli, S. O., Ficner, R., Irniger, S., & Braus, G. H. (2012). The *Aspergillus nidulans* MAPK module AnSte11-Ste50-Ste7-Fus3 controls development and secondary metabolism. *PLoS Genetics*, 8(7), 1-19.
- Benítez, T., Rincón, A. M., Limón, M. C., & Codón, A. C. (2004). Biocontrol mechanisms of *Trichoderma* strains. *International Microbiology: The Official Journal of The Spanish Society for Microbiology*, 7(4), 249–260.
- Biopython. Python Tools for Computational Molecular Biology. **https://biopython.org/**.
- BLAST. (2008). Command Line Applications User Manual. Bethesda (MD): National Center for Biotechnology Information (US). Table C1: [Options common to all BLAST+...]. **https://www.ncbi.nlm.nih.gov/books/NBK279684/table/appendices.T.options_common_to_all_blast/**.
- Brewster, J. L., & Gustin, M. C. (2014). Hog1: 20 years of discovery and impact. *Science Signaling*, 7(343), re7,1-10.
- Brown, G. D., Denning, D. W., Gow, N. A. R., Levitz, S. M., Netea, M. G., & White, T. C. (2012). Hidden Killers: Human Fungal Infections. *Science Translational Medicine*, 4(165), 165rv13–165rv13.
- Cairns, T.C., Nai, C. & Meyer, V. (2018). How a fungus shapes biotechnology: 100 years of *Aspergillus niger* research. *Fungal Biology and Biotechnology*, 5(13), 1-14.
- Calisto, F. M., Vermeulen, W. J. V., & Salomone, R. (2020). A typology of circular economy discourses: Navigating the diverse visions of a contested paradigm. *Resources, Conservation and Recycling*, 161, 104917.
- Chen, C., Hou, J., Tanner, J. J., & Cheng, J. (2020). Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *International Journal of Molecular Sciences*, 21(8), 2873.
- Cheung, N., Tian, L., Liu, X., & Li, X. (2020). The Destructive Fungal Pathogen *Botrytis cinerea*-Insights from Genes Studied with Mutant Analysis. *Pathogens (Basel, Switzerland),* 9(11), 923.
- Chuan, Lin., Young-rae, Cho., Woo-chang, Hwang., Pengjun, Pei., & Aidong, Zhang. (2006). Clustering Methods in Protein-Protein Interaction Network. Department of Computer Science and Engineering, State University of New York at Buffalo. Chapter 1, 1-35.
- Cole, G. T. (1996). Basic Biology of Fungi. In S. Baron (Ed.), *Medical Microbiology*. (4th ed.). University of Texas Medical Branch at Galveston.
- Daisuke, H., Kazutoshi, S., Keietsu, A., & Katsuya, G. (2016). Signaling pathways for stress responses and adaptation in *Aspergillus* species: stress biology in the post-genomic era, *Bioscience, Biotechnology, and Biochemistry*, 80(9), 1667–1680.

- Davies, A., Hooley, F., Causey-Freeman, P., Eleftheriou, I., & Moulton, G. (2020). Using interactive digital notebooks for bioscience and informatics education. *PLoS Computational Biology,* 16(11), e1008326.
- David, H., Ozçelik, I. S., Hofmann, G., & Nielsen, J. (2008). Analysis of Aspergillus nidulans metabolism at the genome-scale. BMC Genomics, 9(163), 1-15.
- de Nadal, E., & Posas, F. (2022). The HOG pathway and the regulation of osmoadaptive responses in yeast. *FEMS Yeast Research*, *22*(1), 1-7.
- de Vries, R. P., Riley, R., Wiebenga, A., Aguilar-Osorio, G., Amillis, S., Uchima, C. A., … Barry, K. (2017). Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus *Aspergillus*. *Genome Biology*, 18(1), 1-45.
- Duran, R., Cary, J. W., & Calvo, A. M. (2010). Role of the osmotic stress regulatory pathway in morphogenesis and secondary metabolism in filamentous fungi. *Toxins*, 2(4), 367–381.
- Ebbole, D. J. (1996). Morphogenesis and vegetative differentiation in filamentous fungi. *Journal of Genetics*, 75(3), 361–374.
- Egbuta, M.A., Mwanza, M., Babalola, O.O. (2016). A Review of the Ubiquity of Ascomycetes Filamentous Fungi in Relation to Their Economic and Medical Importance. *Advances in Microbiology*, 6, 1140-1158.
- Folch, M. J. L., Garay, A. A., Lledías, F., & Covarrubias R. A. A. (2004). La respuesta a estrés en la levadura *Saccharomyces cerevisiae. Revista Latinoamericana de Microbiología,* 46(1-2), 24-46.
- Fones, H., & Gurr, S. (2015). The impact of Septoria tritici Blotch disease on wheat: An EU perspective. *Fungal Genetics and Biology*: FG & B, 79, 3–7.
- Friedman, D. Z. P., & Schwartz, I. S. (2019). Emerging Fungal Infections: New Patients, New Patterns, and New Pathogens. *Journal of Fungi* (Basel, Switzerland), 5(3), 1-19.
- Furukawa, K., Hoshi, Y., Maeda, T., Nakajima, T., & Abe, K. (2005). *Aspergillus nidulans* HOG pathway is activated only by two-component signalling pathway in response to osmotic stress. *Molecular Microbiology*, 56(5), 1246–1261.
- Garrido, B. V., Jaimes, A. R., Sánchez, O., Lara, R. F., & Aguirre, J. (2018). SakA and MpkC Stress MAPKs Show Opposite and Common Functions During Stress Responses and Development in *Aspergillus nidulans*. *Frontiers in Microbiology*, 9(2518), 1-12.
- Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2018). A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6), 1981–1996.
- Gennarelli, M., & Cattaneo, A. (2010). Genetic Variations and Association. Pharmacology of 5-HT6 Receptors. *International Review of Neurobiology*, - Part 1, 129–151.
- Hernández, C. R., Pinto, A. R., Arenas, R., Sánchez, C. C. D., Espinosa, H. V. M., Sierra, M. K. Y., Conde, C. E., Juárez, D. E. R., Xicohtencatl, C. J., Carrillo, C. E. M., Steven, V. J., Martínez, H. E., & Rodríguez, C. C. (2022). Epidemiology of Clinical Sporotrichosis in the Americas in the Last Ten Years. *Journal of Fungi* (Basel, Switzerland), 8(6), 1-29.
- Hernández, D. E. M., Castillo, O. L. S., García, E. Y., Mandujano, G. V., Díaz, G. G. & Álvarez C. J. (2019). Bioinformatics as a Tool for the Structural and Evolutionary Analysis of Proteins. In P. Behzadi, & N. Bernabò (Eds.), Computational Biology and Chemistry. IntechOpen. doi: 10.5772/intechopen.89594.
- Hibbett, D. S., Binder, M., Bischoff, J. F., Blackwell, M., Cannon, P. F., Eriksson, O. E., Huhndorf, S., James, T., Kirk, P. M., Lücking, R., Thorsten Lumbsch, H., Lutzoni, F., Matheny, P. B., McLaughlin, D. J., Powell, M. J., Redhead, S., Schoch, C. L., Spatafora, J. W., Stalpers, J. A., Vilgalys, R., … Zhang, N. (2007). A higher-level phylogenetic classification of the Fungi. Mycological research, 111(Pt 5), 509–547.
- Hohmann, S. & Mager, W. (1997). Shaping up: The responses of yeast to osmotic stress. Yeast stress responses. U.S.A., Chapman & Hall:101-146.
- Huerta, G., Holguín, F., Benítez, F. & Toledo, J. (2009). Epidemiología de la Antracnosis [*Colletotrichum gloeosporioides* (Penz.) Penz. and Sacc.] en Mango (Mangifera indica L.) cv. Ataulfo en el Soconusco, Chiapas, México. *Revista Mexicana de Fitopatología*, 27(2), 93-105.
- Hüttner, S., Johansson, A., Gonçalves T. P., Achterberg P. & Nair B. R. (2020). Recent advances in the intellectual property landscape of filamentous fungi. *Fungal Biology and Biotechnology* 7 (16), 1-17.

- INSST. (2021). Fichas de agentes biológicos-Hongos: *Trichophyton rubrum*. BaseBIO INSST, Madrid España.
- Interreg. Baltic Sea Region, European Union. (2020). Circularity strategies: Enterprises apply different strategies in relation to design, production and recovery. Infographic of circularpp.eu.
- Jiang, R., Zhang, X., & Zhang, M. Q. (Eds.). (2013). Basics of Bioinformatics: Lecture Notes of the Graduate Summer School on Bioinformatics of China, 271-301. doi:10.1007/978-3-642-38951-1.
- Keerthikumar, S. (2017). An Introduction to Proteome Bioinformatics. *Methods in Molecular Biology (Clifton, N.J.),* 1549, 1–3.
- Kendrick, B. (1985). The Fifth Kingdom (3rd Ed.), 1, Mycology Publications, Waterloo, Ont., Canada.
- Kubicek, C. P., Steindorff, A. S., Chenthamara, K., Manganiello, G., Henrissat, B., Zhang, J., Cai, F., Kopchinskiy, A. G., Kubicek, E. M., Kuo, A., Baroncelli, R., Sarrocco, S., Noronha, E. F., Vannacci, G., Shen, Q., Grigoriev, I. V., & Druzhinina, I. S. (2019). Evolution and comparative genomics of the most common *Trichoderma* species. *BMC Genomics*, 20(1), 485, 1-24.
- Kumar, A. (2020). *Aspergillus nidulans*: A Potential Resource of the Production of the Native and Heterologous Enzymes for Industrial Applications. *International Journal of Microbiology*, 2020(8894215), 1-11.
- Li, J., Lin, L., Sun, T., Xu, J., Ji, J., Liu, Q., & Tian, C. (2020). Direct production of commodity chemicals from lignocellulose using *Myceliophthora thermophila*. *Metabolic Engineering*, 61, 416–426.
- Li, Y., Steenwyk, J. L., Chang, Y., Wang, Y., James, T. Y., Stajich, J. E., Spatafora, J. W., Groenewald, M., Dunn, C. W., Hittinger, C. T., Shen, X. X., & Rokas, A. (2021). A genome-scale phylogeny of the kingdom Fungi. *Current Biology*: CB, 31(8), 1653–1665.e5.
- Liu, L., Chen, Z., Liu, W., Ke, X., Tian, X. & Chu, J. (2022). Cephalosporin C biosynthesis and fermentation in *Acremonium chrysogenum*. *Applied Microbiology and Biotechnology,* 106(19-20), 6413–6426.
- Lübeck, M., & Lübeck, P. S. (2022). Fungal Cell Factories for Efficient and Sustainable Production of Proteins and Peptides. *Microorganisms*, 10(4), 753, 1-24.
- Madden, T. (2010). The BLAST Sequence Analysis Tool. The NCBI Handbook. The National Library of Medicine. Chapter 6, 1-15.
- Mager, W. H., de Boer, A. H., Siderius, M. H., & Voss, H. P. (2000). Cellular responses to oxidative and osmotic stress. *Cell Stress & Chaperones,* 5(2), 73–75.
- Mangalam, H. (2002). The Bio* toolkits--a brief overview. Briefings in bioinformatics, 3(3), 296–302.
- Martínez, R., Cortés, C., Madrigal, L. & González, J. (2019). Fungi and Yeasts: Lipase Factories. *Interciencia*, 44(7), 378-385.
- Martínez, V. R., Garza, R. T. S., Moreno, M. V. R., Hernández, D. S., & Mayek, P. N. (2016). Bases bioquímicas de la tolerancia al estrés osmótico en hongos fitopatógenos: el caso de *Macrophomina phaseolina (Tassi) Goid*. *Revista Argentina de Microbiología,* 48(4), 347-357.
- McGinnis, MR, Tyring SK. (1996). Introduction to Mycology. Medical Microbiology. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston.
- McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32 32(Web Server issue), W20–W25.
- Meyer, V., Andersen, M. R., Brakhage, A. A., Braus, G. H., Caddick, M. X., Cairns, T. C., … Head, R. M. (2016). Current challenges of research on filamentous fungi in relation to human welfare and a sustainable bio-economy: a white paper. *Fungal Biology and Biotechnology*, 3(1), 1-17.
- Meyer, V., Basenko, E. Y., Benz, J. P., Braus, G. H., Caddick, M. X., Csukai, M., de Vries, R. P., Endy, D., Frisvad, J. C., Gunde-Cimerman, N., Haarmann, T., Hadar, Y., Hansen, K., Johnson, R. I., Keller, N. P., Kraševec, N., Mortensen, U. H., Perez, R., Ram, A., Record, E., … Wösten, H. (2020). Growing a circular economy with fungal biotechnology: a white paper. *Fungal Biology and Biotechnology,* 7:5, 1-23.
- Miskei, M., Karányi, Z., & Pócsi, I. (2009). Annotation of stress-response proteins in the aspergilli. *Fungal Genetics and Biology*, 46 Suppl 1, S105–S120.

- Mojsov, K. D. (2016). New and Future Developments in Microbial Biotechnology and Bioengineering: *Aspergillus* System Properties and Applications, *Aspergillus* Enzymes for Food Industries. Chapter 16, 215–222.
- Molina, A., Gómez, L. & Umaña, L. (2017). Identificación de especies del género *Colletotrichum* asociadas a la antracnosis en papaya (Carica papaya L.) en Costa Rica. *Agronomía Costarricense,* 41(1), 69-80.
- Muggia, L., Ametrano, C. G., Sterflinger, K., & Tesei, D. (2020). An Overview of Genomics, Phylogenomics and Proteomics Approaches in *Ascomycota*. *Life (Basel, Switzerland)*, 10(12), 356.
- NCBI. BLAST: Basic Local Alignment Search Tool. **https://blast.ncbi.nlm.nih.gov/Blast.cgi**.
- Pandas. (2022). Documentación: Marco de datos: pandas.DataFrame. **https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html**.
- Pathak, V. M., & Navneet. (2017). Review on the current status of polymer degradation: a microbial approach. *Bioresources and Bioprocessing,* 4(1), 1-31.
- Pazos, F., & Chagoyen, M. (2015). Practical Protein Bioinformatics. Springer, ISBN 978-3-319-12726-2. **https://doi.org/10.1007/978-3-319-12727-9.**
- Persson, B. (2000). Bioinformatics in protein analysis. Experientia supplementum (EXS*),* Edition by P. Joll"s and H.Jornvall, 88, 215–231.
- Pessôa, M. G., Paulino, B. N., Mano, M. C. R., Neri-Numa, I. A., Molina, G., & Pastore, G. M. (2017). *Fusarium* species-a promising tool box for industrial biotechnology. *Applied Microbiology and Biotechnology*, 101(9), 3493–3511.
- Project Jupyter. Project Jupyter[Internet]. 2019. En Jupyter. **https://jupyter.org/** .
- Raud, M., Kikas, T., Sippula, O., & Shurpali, N. J. (2019). Potentials and challenges in lignocellulosic biofuel production technology. *Renewable and Sustainable Energy Reviews,* 111, 44–56.
- Rispail, N., Soanes, D. M., Ant, C., Czajkowski, R., Grünler, A., Huguet, R., … Di Pietro, A. (2009). Comparative genomics of MAP kinase and calcium–calcineurin signalling components in plant and human pathogenic fungi. *Fungal Genetics and Biology*, 46(4), 287–298.
- Rocher, F., Alouane, T., Philippe, G., Martin, M. L., Label, P., Langin, T., Bonhomme, L. (2022). *Fusarium graminearum* Infection Strategy in Wheat Involves a Highly Conserved Genetic Program That Controls the Expression of a Core Effectome. *International Journal of Molecular Sciences*. 23(3):1914.
- Rojo, I., Álvarez, B., García, R. S., León, J., Sañudo, A., & Allende, R. (2017). Situación actual de *Colletotrichum* spp. en México: Taxonomía, caracterización, patogénesis y control. *Revista Mexicana de Fitopatología*, 35(3), 549-570.
- Rokas, A. (2022). Evolution of the human pathogenic lifestyle in fungi. *Natural Microbiology* 7, 607–619.
- Różewicz, M., Wyzińska, M., Grabiński, J. (2021). The Most Important Fungal Diseases of Cereals—Problems and Possible Solutions. *Agronomy*, 11(4), 1-12.
- Santamaría, F., Díaz, R., Gutiérrez, O., Santamaría, J. & Larqué, A. (2011). Control de dos especies de *Colletotrichum* causantes de antracnosis en frutos de papaya Maradol. *Revista Mexicana de Ciencias Agrícolas*, 2(5), 631-643.
- Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, 3(3), 430–439.
- Scholz, M. (2022). Metagenomics: Computational tools. **https://www.metagenomics.wiki/tools/blast/blastn-output-format-6**.
- Sharma, M. & Kulshrestha, S. (2015). *Colletotrichum gloeosporioides*: An anthracnose causing pathogen of fruits and vegetables. *Biosciences Biotechnology Research Asia,* 12(2), 1233-1246.
- Sharmeen, N., Sulea, T., Whiteway, M., & Wu, C. (2019). The adaptor protein Ste50 directly modulates yeast MAPK signaling specificity through differential connections of its RA domain. *Molecular Biology of The Cell*, 30(6), 794–807.
- Singh, B. (2014). *Myceliophthora thermophila. Sporotrichum thermophile*: a thermophilic mould of biotechnological potential. *Critical Reviews in Biotechnology*, 36(1), 59–69.
- Skoneczny, M. (2018). Stress Response Mechanisms in Fungi. Springer Cham. Edition 1. ISBN: 978-3-030-00683-9. **https://doi.org/10.1007/978-3-030-00683-9**.

- Soltani, J. (2016). New and Future Developments in Microbial Biotechnology and Bioengineering: *Aspergillus* System Properties and Applications. Secondary Metabolite Diversity of the Genus *Aspergillus:* Recent Advances., Chapter 22, 275–292.
- Spatafora, J. W., Aime, M. C., Grigoriev, I. V., Martin, F., Stajich, J. E., & Blackwell, M. (2017). The Fungal Tree of Life: from Molecular Systematics to Genome-Scale Phylogenies. *Microbiology Spectrum*, 5(5), 1-32.
- Surabhi, P. & Onkar, S. (2020). Cephalosporin Market by Generation (First-generation, Second-generation, Third-Generation, Fourth-Generation, and Fifth-Generation), Type (Branded and Generic), Route of Drug Administration (Intravenous and Oral), and Application (Respiratory Tract Infection, Skin Infection, Ear Infection, Urinary Tract Infection, and Sexually Transmitted Infection): Global Opportunity Analysis and Industry Forecast, 2019–2027. Report Code: A03170. Allied Market Research.
- Tapia, R. A., Ramírez, D. J. F., Salgado, S. M. L., Castañeda V. Á., Maldonado, Z. F. I., & Lara, D. A. V. (2020). Distribución espacial de antracnosis (*Colletotrichum gloeosporioides* Penz) en aguacate en el Estado de México, México. *Revista Argentina de Microbiología,* 55(1), 72-81.
- Thambugala, K. M., Daranagama, D. A., Phillips, A. J. L., Kannangara, S. D., & Promputtha, I. (2020). Fungi vs. Fungi in Biocontrol: An Overview of Fungal Antagonists Applied Against Fungal Plant Pathogens. *Frontiers in Cellular and Infection Microbiology*, 10(604923), 1-19.
- Torrades, O. S. (2004). Proteómica: El diseño molecular de la vida. *Offarm*, 23(4), 126-130.
- Virag, A., Lee, M. P., Si, H., & Harris, S. D. (2007). Regulation of hyphal morphogenesis by cdc42 and rac1 homologues in *Aspergillus nidulans*. *Molecular Microbiology,* 66(6), 1579–1596.
- Wang, Q., Zhong, C., & Xiao, H. (2020). Genetic Engineering of Filamentous Fungi for Efficient Protein Expression and Secretion. *Frontiers in Bioengineering and Biotechnology*, 8(293), 1-8.
- Wilson, W., Dahl, B., Nganje, W. (2018). Economic Costs of Fusarium Head Blight, Scab and Deoxynivalenol. *World Mycotoxin Journal,* 11(2), 291–302.
- Yang, L., Lübeck, M., & Lübeck, P. S. (2017). *Aspergillus* as a versatile cell factory for organic acid production. *Fungal Biology Reviews*, 31(1), 33–49.

Vo. Bo. DE LOS ASESORES

Dr. JUAN ESTEBAN BARRANCO FLORIDO

Dr. JESÚS EDUARDO ZÚÑIGA LEÓN

Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA UNIDAD XOCHIMILCO

DIVISIÓN DE CIENCIAS BIOLÓGICAS Y DE LA SALUD

DEPARTAMENTO DE SISTEMAS BIOLÓGICOS

LICENCIATURA EN QUÍMICA FARMACEÚTICA BIOLÓGICA

INFORME DE ACTIVIDADES DEL SERVICIO SOCIAL:

Caracterización *in silico* de la vía de señalización de respuesta a estrés osmótico en hongos filamentosos de la división *Ascomycota*.

PROYECTO GENÉRICO:

Obtención de materias primas, principios activos, medicamentos y productos biológicos.

PRESENTA:

Zarza Sánchez Larissa

Matrícula: 2173082232

Tutores:

Dr. Juan Esteban Barranco Florido No. Eco. 24927

Dr. Jesús Eduardo Zúñiga León Cédula Profesional 10977107

LUGAR DE REALIZACIÓN: Laboratorio De Biotecnología, Edificio N. Departamento de Sistemas Biológicos, Universidad Autónoma Metropolitana Unidad Xochimilco, con Dirección: Calzada Del Hueso 1100. Col. Villa Quietud, Alcaldía de Coyoacán, C.P 04960, Ciudad de México, México.

Periodo: 17 de noviembre de 2021 al 17 de mayo de 2022

Septiembre 2023

9. Resumen

Los hongos filamentosos juegan un papel fundamental en áreas relacionadas al ser humano como lo es la industria alimenticia, farmacéutica, agrícola, médica, papelera e investigación científica. Muchos de ellos pertenecen a la división *Ascomycota*, en la cual encontramos hongos patógenos de plantas, insectos y humanos. Estos poseen vías intracelulares de señalización que les permiten responder a diferentes tipos de estrés, sin embargo, aunque la mayoría de los hongos comparten mecanismos conservados, estos no son idénticos, por lo que es importante su optima identificación. El estrés osmótico es uno de los mecanismos de respuesta más estudiados en el hongo *Saccharomyces cerevisiae*, sin embargo, de los hongos filamentosos los más estudiados son algunas especies de *Aspergillus*, como lo es el hongo modelo *Aspergillus nidulans/Emericella nidulans*.

La proteómica y las herramientas bioinformáticas son indispensables en la identificación y caracterización de proteínas, relaciones funcionales, así como en la evolución de la dinámica celular. Los lenguajes de programación se han acoplado a esto con el fin de poder procesar datos a gran escala y facilitar el análisis.

En este proyecto se identificaron proteínas ortólogas de la vía de respuesta a estrés osmótico en hongos filamentosos de la división *Ascomycota,* utilizando la interfaz Jupyter notebook, el lenguaje Python y bases de datos. La identificación de ortólogos se hizo mediante el programa BLAST y con ayuda de Phobius y Pfam se captaron las proteínas de mayor similitud dada su arquitectura. Después por el índice de jaccard se tomaron en cuenta solo aquellas con el mismo número de dominio(s) idénticos o casi idénticos. Posteriormente para buscar el mejor hit para cada proteína se hizo una búsqueda BLAST reciproca de All vs All de los datos obtenidos. Finalmente, por el análisis de clustering se visualizaron clústeres de proteínas potencialmente ortólogas y ante la construcción de un Heatmap se visualizó el grado de conservación de estas proteínas y por tanto de la vía de respuesta a estrés osmótico en hongos ascomicetos, permitiendo comparar su dinámica molecular. Como resultado se identificaron 191 proteínas potencialmente ortólogas implicadas en la vía de señalización. Al ejecutar el Análisis de Clustering se formaron 15 grupos o clústeres de proteínas con alta similaridad. El género *Trichoderma* ha destacado en el estudio *in silico* debido a la conservación de arquitectura de dominio en proteínas quinasas. La interfaz Jupyter notebook y el lenguaje informático Python son herramientas útiles para jóvenes investigadores y han favorecido el estudio *in sillico* de proteínas.

10. Anexos

Anexo 1. Librerías importadas a Python: Pandas, NumPy, Matplotlib, Seaborn, SciPy.

```
import pandas as pd
import numpy as np
from pandas import DataFrame

import matplotlib.pyplot as plt
import seaborn as sns
from scipy.cluster.hierarchy import dendrogram, linkage
import matplotlib as mpl
```

Anexo 2. Código Phobius y Pfam.

```
cols = ['Entry', 'alignment start', 'alignment end', 'start', 'end', 'hmm acc',
        'name', 'type', 'hmm start', 'hmm end', 'hmm length', 'bit score', 'E-value', 'significance','clan']
```

```
res0 = []
with open('pfam_sequences_Larissa.txt', 'r') as fq:
    for line in fq:
        line = line.rstrip()
        if '#' in line:
            pass
        else:
            line = re.sub(' * ', ' ', line)
            res0.append(line.split(' ')[0:15])
```

```
hmm_all = DataFrame(res0, columns = cols).dropna()
hmm_all['E-value'] = hmm_all['E-value'].astype(float)
```

```
pfam=hmm_all[["Entry","start","end","name","type",]]
```

```
tc=[]
with open ("phobius_sequences_Larissa.txt","r") as ph:
    for i in ph:
        i = i.rstrip()
        if "##" in i:
            s=[]
        else:
            s.append(i)
        if len(s)>1:
            if ("SIGNAL" in "".join(s)) or ("TRANSMEM" in "".join(s)):
                ide=s[0].replace("ID   ","")
                for i in s[1:]:
                    if "SIGNAL" in i:
                        esp=re.sub(" * ","\t",i)
                        ps=esp.split("\t")
                        #print(ide,ps[2],ps[3],ps[1])
                        tc.append([ide,ps[2],ps[3],"sig_p","Signal_peptide"])
                    if "TRANSMEM" in i:
                        esp1=re.sub(" * ","\t",i)
                        sp=esp1.split("\t")
                        #print(ide,sp[2],sp[3],sp[1])
                        #identificador, inicio, final y tipo
                        tc.append([ide,sp[2],sp[3],"transmembrane","Transmembrane"])
```

```
fobius=DataFrame(tc,columns=["Entry","start","end","name","type"]).drop_duplicates().reset_index(drop = True)
fobius
```

Anexo 3. Código para la ejecución del análisis de clustering y su representación. (dendograma).

```
# estas son las métricas que soporta el análisis de clustering
metricas = ['euclidean','braycurtis','canberra','chebyshev','cityblock','correlation','cosine','dice',
            'hamming','jaccard','jensenshannon','kulsinski','mahalanobis',
            'matching','minkowski','rogerstanimoto','russellrao','seuclidean','sokalmichener',
            'sokalsneath','sqeuclidean','yule']
```

```
clusters = linkage(matmat.values.T, method = 'complete', metric='correlation')
# se realiazó un análisis de clustering usando la métrica de correlación ya que esta permitió
# identificar más grupos de proteínas ortólogas en múltiples organismos
```

```
mpl.rcParams.update(mpl.rcParamsDefault)
fig, ax = plt.subplots(figsize=(5, 20))

dend2 = dendrogram(clusters,
            orientation='left',truncate_mode='level',
            labels=matmat.columns)
#ax.axis('off')

ax.tick_params(axis='y', which='major', labelsize=6)
#plt.savefig('dendrogram.png', dpi = 900, bbox_inches='tight')
plt.show()
```

Anexo 4. Código para crear el *heatmap*.

```
from colormap import Colormap
```

```
sns.set(font_scale=0.2)

ax = sns.heatmap(mat1.values, linewidth = 0,
                 xticklabels= mat.columns.tolist()[1:],
                 yticklabels= mat.organism_sacc.tolist(),
                 cmap = Colormap().cmap_linear('#eaeaf2', 'orange', 'black'))
#plt.savefig('all_vs_all.png', dpi = 1920, bbox_inches='tight')

#plt.scatter([mat1.index.tolist()[-1]] * len(mat1), mat1.index.tolist(), s = 0.5, c = 'black')

plt.show()
```

```
import warnings
warnings.filterwarnings("ignore")
```

```
sns.set(font_scale=0.2)
g = sns.clustermap(mat1.values.T, method = 'complete', metric="correlation",
                   tree_kws=dict(colors=dend2['color_list']),
                   linewidths=0, linecolor='white',
                   square=True, col_cluster = False, row_cluster = True,
                   xticklabels= mat.columns.tolist()[1:],
                   yticklabels= mat.organism_sacc.tolist(),
                   #mask = mat1.values <= float(50),
                   #annot = True, annot_kws={"size": 5},fmt= '.2g',
                   cbar_pos=(1, 0.1, .03, .19),
                   cmap = Colormap().cmap_linear('#eaeaf2', 'orange', 'black'), figsize = (10,10))

for a in g.ax_row_dendrogram.collections:
    a.set_linewidth(1.5)
for a in g.ax_col_dendrogram.collections:
    a.set_linewidth(1.5)

g.ax_heatmap.tick_params(bottom=True, right=True, top=False, left=False, width = 0.2, length=4, color='black')

plt.savefig('stress_response.png', dpi = 900, bbox_inches='tight')
plt.show()
```

Vo. Bo. DE LOS ASESORES

Dr. JUAN ESTEBAN BARRANCO FLORIDO

Dr. JESUS EDUARDO ZÚÑIGA LEÓN